DOCUMENT RESUME

ED 070 084 CS 200 253

AUTHOR	Hoetker, James
TITLE	Students as Audiences: An Experimental Study of the Relationships between Classroom Study of Drama and Attendance at the Theatre. Research Report No. 11.
INSTITUTION	Central Midwestern Regional Educational Lab., St. Ann, Mo.; National Council of Teachers of English, Champaign, Ill.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-6-2875
PUB DATE	71
CONTRACT	OEC-3-7-062875-3056
NOTE	143p.
AVAILABLE FROM	National Council of Teachers of English, 1111 Kenyon Road, Urbana, Ill. 61801 (Stock No. 16057, $2.75 non-member, $2.50 member)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Content Analysis; *Drama; Dramatics; Educational Objectives; Educational Research; *English Instruction; Literary Analysis; Literary Conventions; Literary Discrimination; *Literature Appreciation; Research Criteria; *Research Design; Research Methodology; Secondary Education; Teacher Attitudes; *Teaching Methods

ABSTRACT

This study is a revision of a previous study, sponsored by CEMREL, Inc. (see ED 041 903), whose principal purpose was to investigate methods of teaching dramatic literature; additional information was obtained on the practical problems involved in educational research and the utilization of some basic aspects of multivariate fractional factorial design. Disagreement between English teachers and professional theatre personnel about the best methods of preparing students to attend dramatic productions led to a 6-month study that involved 52 teachers and more than 1,300 students. Study materials were supplied to students in conjunction with their classroom discussions and the viewing of two plays. Independent variables considered for each play were (1) intensity of study of the background, (2) intensity of study of the text, (3) timing of the classroom treatment, (4) content of the classroom treatment. Fifteen dependent variables were used. Generally, the relatively few significant effects confirmed the supposition that English teachers preferred arrangements yielding the highest scores on the cognitive tasks; and the actors preferred arrangements maximizing scores in appreciation and affective response. Extensive discussion is presented on the research design, and appendices include sample tests and tables of results. (Author/DI)

ED 070084

NCTE Research Report No. 11

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Students *as Audiences:

*An Experimental Study of the Relationships between Classroom Study of Drama and Attendance at the Theatre

JAMES HOETKER
University of Illinois

National Council of Teachers of English
508 South Sixth Street, Champaign, Illinois 61820

CS 200 253

NATIONAL COUNCIL OF TEACHERS OF ENGLISH

Committee on Research

James R. Squire, Ginn and Company, *Chairman*
Doris V. Gunderson, U.S. Office of Education, *Associate Chairman*
Thomas C. Barrett, University of Wisconsin/Nathan S. Blount, University of Wisconsin/John R. Bormuth, University of Chicago/Earl W. Buxton, University of Alberta/James Hoetker, University of Illinois/Stanley B. Kegler, University of Minnesota/Geraldine E. LaRocque, Teachers College, Columbia University/Evan Glyn Lewis, University of Wales/Roy C. O'Donnell, Florida State University/Walter T. Petty, State University of New York at Buffalo/Alan C. Purves, University of Illinois/Mildred E. Riling, Southeastern State College, Oklahoma/William D. Sheldon, Syracuse University/J. Stephen Sherwin, State University of New York at Buffalo/Robert F. Hogan, NCTE, *ex officio*/Bernard O'Donnell, NCTE/ERIC, *ex officio*

Consultant Readers for This Manuscript

John R. Bormuth, University of Chicago/Geraldine E. LaRocque, Teachers College, Columbia University/Alan C. Purves, University of Illinois

Committee on Publications

Robert F. Hogan, NCTE Executive Secretary, *Chairman*/Robert Dykstra, University of Minnesota/Walker Gibson, University of Massachusetts, Amherst/Mildred E. Webster, St. Joseph Senior High School, Michigan/Eugene C. Ross, NCTE Director of Publications

Editorial Services Robyn Grant D'Alba, NCTE Headquarters
Design Norma Phillips Meyers, NCTE Headquarters

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Library of Congress Catalog Card Number 79-140904
ISBN 0-8141-1605-7
NCTE Stock Number 16057

Copyright © 1971 by the National Council of Teachers of English
508 South Sixth Street, Champaign, Illinois 61820
Printed in U.S.A.

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY National Council of Teachers of English TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

National Council of Teachers of English
Research Report No. 11

Many studies of the past have attempted to study both affective and cognitive dimensions of student response to literature. Few have suggested so clearly the complex interaction of these two modes of reader response as does the present study by James Hoetker. Moreover, beyond suggesting the complexity and diversity of pupil responses, Hoetker demonstrates that the approach emphasized in instruction can affect in diverse ways the reactions of young people to dramatic literature, albeit not so clearly as partisans arguing for different emphases would like to believe.

The Committee on Research is pleased to introduce this study not only for the significance of its findings but for the sharpness of its design and the clarity of its statistical presentation. The future of empirical research in the teaching of English is dependent no less upon the importance of asking important researchable questions than upon utilizing empirical methods appropriate to the basic assumptions of a study. In this study James Hoetker clearly related question and design in a manner that yields findings in which the profession can have considerable confidence.

James R. Squire
For the Committee on Research

TABLE OF CONTENTS

RESEARCH REPORT NO. 11 v

TABLES AND SUMMARY TABLES ix

ACKNOWLEDGMENTS xi

INTRODUCTION xiii

CHAPTER ONE. DEFINING THE PROBLEM 1

- The Positions to Be Evaluated 2
- The Objectives-for-Drama Study 6

CHAPTER TWO. THE EXPERIMENTAL DESIGN 9

CHAPTER THREE. DEFINING THE INDEPENDENT VARIABLES 27

- Timing 28
- Plays 28
- Content 29

CHAPTER FOUR. THE DEPENDENT VARIABLES 33

- The Objectives-for-Drama Categories 33
- Discussion of the Dependent Variables 35
- The Validity of the Instruments: Some Comments 44
- Other Measures 47

CHAPTER FIVE. THE PLAN FOR ANALYZING THE DATA 51

- The Contrasts 51
- Analyses of Effects and Interactions 61

CHAPTER SIX. OTHER FEATURES OF THE STUDY ... 65
Item Sampling ... 65
Assignment of Subjects to Treatments ... 67
Three Uncontrolled Sources of Variation ... 67

CHAPTER SEVEN. PRESENTATION OF RESULTS ... 73
Summary of Significant Contrasts ... 73
Significant Findings Under Each Hypothesis ... 76
Other Findings ... 87

CHAPTER EIGHT. CONCLUSIONS ... 89
Summary of Significant Effects ... 89
Comments ... 89

APPENDIX A. DISTRIBUTION OF TEST ITEMS OVER FORMS ... 95

APPENDIX B. ADDITIONAL OBSERVATIONS ON THE CONDUCTING OF THE STUDY ... 117

APPENDIX C. SPECIFICATIONS OF TREATMENTS AND MATERIALS FOR DESIGN CONDITION #4 ... 125
First Play: O'Casey's *Red Roses for Me* ... 125
Second Play: Shakespeare's *Macbeth* ... 126

APPENDIX D. SUMMARY TABLES OF F-RATIOS AND SIGNIFICANCE LEVELS FOR TOTAL SCORES AND CATEGORIES UNDER ALL HYPOTHESES ... 127

TABLES AND SUMMARY TABLES

Table 1	A 2^2 Factorial Design	15
Table 2	A 2^3 Factorial Design Matrix	16
Table 3	Summary of Variables and Levels of Variables in the Experimental Study	17
Table 4	Design Matrix for a 2^{5-1} Fractional Factorial Design	18
Table 5	The Design for the Experimental Teaching Study (First Version)	21
Table 6	The Design Matrix for the Experimental Teaching Study (Second Version)	25
Table 7	Titles and Code Designations of All Dependent Variables	35
Table 8	Mean Total Scores on All Dependent Measures at Two Levels of an Independent Variable (Fictitious Data for Illustrative Purposes)	52
Table 9	Structure of Contrasts for the MANOVA with the Eleven Student Performance Measures (Total Scores)	53
Table 10	Structure of Contrasts for the MANOVA with the Four Tests in the Affective-Response Category	54
Table 11	Structure of Contrasts for the MANOVA with the Four Tests in the Knowledge Category	54
Table 12	Structure of Contrasts for the MANOVA with the Two Tests in the Philosophical-Understandings Category	55
Table 13	Structure of Contrasts for the MANOVA with the Six Tests in the Appreciation Category	56
Table 14	Structure of Contrasts for the MANOVA with the Six Tests in the Desirable-Attitudes-and-Behaviors Category	57
Table 15	Structure of Contrasts for the MANOVA with the Two Tests in the Interpretive-Skills Category	58

Table 16 Summary of the Hypotheses Evaluated, Plus Other Possible Contrasts and the Alias Structure 62

Table 17 Summary of Significant One- and Two-Factor Effects on Total Test Scores 74

Table 18 Summary of Significant One- and Two-Factor Effects Upon Test Scores within Categories 75

Table 19 Distribution of Test Items over the Three Instruments 96

Summary Table 1A 128
Summary Table 1B 128
Summary Table 2A 129
Summary Table 2B 129
Summary Table 3A 130
Summary Table 3B 130
Summary Table 4A 131
Summary Table 4B 131
Summary Table 5A 132
Summary Table 5B 132
Summary Table 6A 133
Summary Table 6B 133
Summary Table 7A 134
Summary Table 7B 134
Summary Table 8A 135
Summary Table 8B 135
Summary Table 9A 136
Summary Table 9B 136
Summary Table 10A 137
Summary Table 10B 137
Summary Table 11A 138
Summary Table 11B 138

ACKNOWLEDGMENTS

The suggestion that this experimental study be undertaken was originally made by Dr. Wade Robinson, the Executive Director of the Central Midwestern Regional Educational Laboratory (CEMREL). During the planning of the study and through the various stages of preparation for it, Mr. Alan Engelsman and I were about equally responsible for what was done; and Mr. Engelsman continued to be of invaluable assistance throughout the course of the study, although his major responsibilities lay elsewhere.

There were numerous other people without whom the study could never have been successfully completed. Professors David Wiley of the University of Chicago and Tom Johnson of Washington University were primarily responsible for designing the plan for collecting and analyzing the data. Mrs. Charlotte von Breton and Mrs. Lee McClarran were responsible for the day-to-day administration of the study, and expertly supervised the scheduling of the experimental treatments and the physical collection of the raw data. Mr. Saul Hopper assisted valuably in the preparation of the various instruments.

Mr. Richard Robb supervised the handling of the data and the analyses of them. It is impossible to give too much credit to his perseverance and ingenuity.

A large number of Rhode Islanders contributed their time, influence, and advice throughout the project; we are especially grateful in this regard to Miss Rose Vallely, Mr. Donald Rock, Mr. Donald Gardner, and Mr. Richard Cumming. Finally, and perhaps most importantly, we are indebted to the more than fifty teachers who bore with us throughout the study, to their students, and to the English department chairmen and the principals of the fourteen Rhode Island schools which participated in the study.

James Hoetker

Champaign, Illinois

INTRODUCTION

This paper attempts to serve three purposes at once. First, it reports the results of an experimental study of the effects of different methods of teaching dramatic literature. Second, it is a case study which should have practical value to anyone who anticipates becoming involved in research in the schools. The experiment itself took more than six months to complete and involved fifty-two teachers and more than 1300 students in fourteen different school districts. Further, it was a remarkably intrusive study, and the teachers involved had to disarrange an entire year's work in order to participate. Despite this, the study went almost precisely as planned, from the administration of pretests in September to the administration of a follow-up test the following April. We have tried to identify, in the course of the reporting, the factors that account for this study's having gone smoothly.

Third, the paper is an introduction in the simplest, most nontechnical language possible to some basic aspects of multivariate factorial experiments. Some of the now quite common techniques used in this study are especially well suited for studying certain areas of English. But such techniques have previously been used by very few researchers in the field. All aspects of this experiment are discussed in plain English in order to introduce these and other techniques to researchers in the language arts who might not be aware of them.

One obvious reason for the poor state of knowledge about research methodology not only in English, but among humanistic educators in general at all levels, is that even the basic texts on experimental design are too technical to be read by most educators. I hope that my own experiences—as an English teacher who has been forced by circumstances to learn something about research—enable me to inform the reader about some of the newer experimental tools that are available, should he wish to make use of them.

Let me make quite clear at the start, though, that I make no claim to any expertise in experimental design, and, if challenged, I will admit to being something of a mathematical illiterate. (Still, I must take full responsibility for any errors in this presentation.) The designing of experiments is a scholarly

specialty, just like Anglo-Saxon literature or modern poetry. It makes no more sense for an English educator to design an experiment than for a statistician to plan a graduate course in English.[1] There is a good deal to be said for the proposition that, with the present state of our knowledge, language arts researchers should be concerned with problems of measurement and theory-building rather than experiments; but, when the question at issue is clearly enough stated that an experiment is called for, the experiment should be a good one. What this means is that one of the first expenditures should be for the services of a specialist in research design.

Let me go a step further and insist that, given our desperate need for empirically-based knowledge, inadequate research studies—i.e., ones designed by amateurs—should no longer be encouraged at any level. Since so much of the published research in English is done by doctoral candidates, it is particularly vital that we stop miseducating doctoral students in English education and similar areas by encouraging them to design their own experimental studies. Individualism in empirical research is a maladaptive anachronism.

[1] I am speaking, of course, of formal studies, intended for publication and involving the expenditure of considerable time and money, not of informal studies designed to give feedback about an ongoing program.

CHAPTER ONE

DEFINING THE PROBLEM

This experiment took place during the third year of our assessment of the Educational Laboratory Theatre Project. This Project involved several federal agencies in cooperatively subsidizing three professional repertory companies so that they might present performances of classic plays to high school students. Three or four plays per year were presented in the three Project sites: Los Angeles, the New Orleans metropolitan area, and the state of Rhode Island. Annually, during the three years of the Project, about one hundred thousand students and teachers were provided experiences with professional theatre that otherwise very few of them would have had.

A feature of the Project was that the primary responsibility for relating the theatre to the school curriculum was given to the English departments of the participating schools. Although the three sites were very different and each had its unique problems and advantages, there were several problems that were common to all the sites. Among the most important of these was that the English teachers and the theatre people had difficulty in understanding one another. The root of the problem was that the two groups had different objectives for the Project, and even when they agreed about objectives, they had different priorities among the common objectives. Especially troublesome were incompatible ideas about the nature and purpose of drama itself.

These disagreements manifested themselves most clearly in disputes about play selection and about the nature and extent of the attention that should be given to the plays in the classroom. The seriousness of the effects of these disagreements upon the operation of the Project in a particular site depended upon the willingness of the school and the theatre representatives to listen to and to learn from one another. But beyond the Project itself, which by now is only of historical interest, the communications problems that characterized the Project have important implications for educators in at least two areas.

First, proposals for using creative and performing artists in various sorts of humanities programs have usually not been very

realistic in assessing the difficulties that may be involved in getting educators and working artists to cooperate. Any program involving such collaboration will involve communications problems due to the kinds of preconceptions that are evaluated in this experiment.

Second, the opinions about drama and literature teaching to which the English teachers in all three sites overwhelmingly subscribed seem to have been learned and profession-specific. The rejection of these opinions by people who are devoting their lives to being exponents and interpreters of dramatic literature cannot be lightly dismissed. To the extent that this experiment is an evaluation of the merits of competing theories about how dramatic literature should be taught, it is of importance to everyone involved in teaching literature, writing literature curricula, and training teachers of English.

The Positions to Be Evaluated

The variations in teaching methods that are examined in this experiment are those that were most prominent in disputes between educators and theatre people. The whole Project was based on the assumptions that appropriate classroom study of the plays would maximize the benefits of the theatre experience and that the availability of a professional performance of a play would enliven and enrich the classroom study of it. Funds had been provided for the preparation of curriculum portfolios to accompany each play; these portfolios, which were distributed to English teachers prior to each performance, contained lesson plans, bulletin board displays, a rich collection of biographical, critical, and historical essays, and various other supplementary materials.[1] Many school administrations had laid down the policy that these portfolios were to be used in each English class before the students attended the plays.

Ironically, the fact that both the educators and the theatre people agreed that what went on in classrooms was vitally important served to heighten the disagreements about how, or whether, the plays should be taught. If the theatre people had

[1] The portfolios or study packets were a regular feature of the Project. In Rhode Island, the portfolios, whose contents were used to define the play-specific classroom treatments, were jointly authored by Rose Vallely, the Project Coordinator for the schools, and Richard Cumming, Trinity Square Repertory Company's Composer-in-Residence and educational officer.

thought that classroom instruction was more or less irrelevant to the students' reception of the performance itself, they would not have cared what went on in the schools. If the educators had not thought that the study of the plays was essential to giving students the full benefit of the performance, they would not have been so concerned about their general lack of special knowledge of theatre or about finding the time to include three or four additional literary works in an already overcrowded curriculum.

All parties to the Project thought that classroom instruction was of vital importance, but educators and theatre people disagreed about what this classroom instruction should include, about how intensive it should be, and about when it should take place. Probably the most clear-cut disagreement was on the matter of the timing of classroom instruction. English teachers and other educators generally advocated classroom study of a play before the performance, so that the students would understand what was going on and therefore be able to enjoy and appreciate the performance. Conversely, most theatre people believed that classroom study should take place only after the production had been seen, with some exceptions to be made as in the case of Shakespeare and other difficult playwrights. The reasons for this difference are fairly clear. The training of the teachers was such that they gave primacy to the literary text of the play and tended to think of the production as an illustration of the text—sort of a super audiovisual device. The following may stand as an extreme statement of the position held by many English teachers.

> Though we must certainly agree that seeing a play and then reading it is better than seeing it and never reading it, we must insist also that to see a play of Shakespeare's before reading it is to damage the experience of reading it. To see one play and then to read a different one is good, and to read the play and thereafter to see it is even better—in fact it is best of all. But to see the play and then to read it is not even as good as merely to read it.[3]

On the other hand, the actors and the directors, thinking of a play as existing essentially only in performance, simply could not see how students could be expected to benefit from talking

[3] Reprinted with permission from *Teaching Shakespeare in the High School* by Bertrand Evans, published by The Macmillan Company. Copyright © 1966. (p. 80)

about a play they had not seen. But the actors also had a more practical reason for wishing the classroom study of the play to come after the performance. Their own experiences with education had convinced many of them that the English teachers would concentrate so wholly on the cognitive aspects of the play and upon right answers that the classroom instruction would interfere with the student's spontaneous affective reactions to the performance. A few of the theatre people were quite vocal in their belief that the teachers would destroy anything they touched and somehow render the play performance as dull as the rest of school. As the director of one of the companies wrote:

> Much if not all of what has been done in school to prepare students for plays has been damaging, I feel, to the excitement and first-time experience of the theatre. . . . Reading a play ahead of time is false; all authors expected their audiences to be experiencing their version of the story for the first time. Few teachers are qualified to excite and lead classes in appreciation for plays, and a pedantic conversion of plot and construction into test material certainly does no good. We have also found that teachers have created improper expectations. . . . I know that it takes longer to awaken the students to what we are actually doing on the stage than if they had had no preparation at all.

> Student audience response has never been bad; and it probably is true that the bad teaching is so bad it simply makes no impressions. . . . The deadliness of the classroom teaching and the compulsory nature of attendance along with forced discussion and examination based on the plays, has for the majority of the students carefully leveled the theatre experience off so that it is safely compatible with the other nonsense which goes on in high school.

In general, the case seemed to be that the theatre people had a great deal more faith in students than the educators did. Teachers thought the students had to be prepared for the theatre; actors thought that the students would respond appropriately if only they were left alone—provided the production was well done. The teachers thought that students had to be taught things so that they could understand plays; the actors thought the plays themselves could teach things. The important point, however, is that everyone agreed that the timing of the study of a play made a significant difference.

It is notable, though, that neither school of thought had any-

thing except personal opinion to support its contentions about this or any other matter. One relevant piece of testimony did rather ambiguously support the English teachers' position. In their *In Search of an Audience*, Brad Morison and Kay Fliehr made the following remarks about student audiences:

> The differences among the reactions of those first student audiences seemed to have little to do with any differences in where the students came from, or with the socioeconomic differences among the high schools. We began to talk with teachers and students at intermission and to listen carefully to the nature of the questions asked after the performance. One difference soon became evident. The more carefully the teachers had prepared the students, the more attentive, well-disciplined, aware, and perceptive they were in the theatre. When the students came from classes where enthusiastic teachers had taught the play well and given them proper perspective on their coming adventure in living theatre, the audiences were enthusiastic. When the students came primarily from classes where the play had only been touched upon in a pedantic manner and the teacher looked upon the trip only as another chaperoning job, the audiences were more restless, less responsive. Apparently the teacher was a very important element in the student's enjoyment of the theatrical experience.[3]

However—and this leads us into discussion of the proper content of the lessons—Morison and Fliehr, when they told about a teacher who did a "thorough and imaginative job of preparing his classes to see *Hamlet*," described a sort of preparation different from that advocated by most English teachers. The teacher Morison and Fliehr used as an example "had chosen *not* to have his classes read the play, but, instead, explored Shakespeare in great detail—his world and his theatre." [4]

This suggestion—that instead of studying the play being performed, students should study everything except the play—had first been voiced by the director of one of the repertory companies. His reasoning was that such a course of study could prime the students to respond to the play, while not depriving them of the pleasures of spontaneous response to it. The same suggestion was later made by other theatre people, and in the passage quoted above at least one English educator finds merit

[3] From *In Search of an Audience* by Bradley G. Morison and Kay Fliehr. Copyright © 1968 by Associated Councils of the Arts. Reprinted by permission of Pitman Publishing Corporation. (p. 192)

[4] Morison, *In Search of an Audience*, p. 193.

in the idea of seeing one play and reading a different one. Typically, though, the English teachers advocated study of the play that was to be staged.

The third matter that everyone agreed was important was the intensity or duration of the classroom study of the play. How much study would get the best results? The English teachers, and most school administrators, believed that a thorough study of the play and its background was essential. The actors tended to think that the less that was done, the better. This matter of intensity was of great practical interest to teachers. They wanted to do all that was necessary, but they found it was impossible to do a thorough study of three or four plays without omitting or slighting other parts of the curriculum. Some protested that an adequate study of each of the plays might end up hurting students who would thereby be given less instruction in those areas included on achievement tests and college entrance examinations.

In summary, then, the experiment being reported here was designed to test a series of theories held by different groups of people involved in a school-theatre project, about the effects of different methods of studying plays. These theories were most importantly concerned with variations in the timing, content, and intensity of the classroom instruction.

The Objectives-for-Drama Study

Some months before we began in earnest to design the experimental study of methods of teaching drama, we undertook a questionnaire study designed to describe the differences between various groups in the objectives that they held for the study of drama in the secondary English class. The study began with the collection of several hundred statements of objectives for drama from a wide variety of printed sources. The objectives were divided on the basis of content analyses into eight categories. Four items from each of these categories were chosen at random and a questionnaire of thirty-two items was made up. Each respondent was to express the strength of his agreement or disagreement with each item on a seven-point scale.

The instrument was administered to samples of English teachers, drama teachers, school principals, and repertory company actors in all three Project sites. The primary finding of this study

was that the four participating groups differed in their objectives for drama as a function of their professional identification.[5]

This study contributed to the experimental study in the following ways. First, factor analyses of the responses to the objectives-for-drama questionnaire helped us to clarify and simplify the categories of objectives we would want to measure in the experimental study. Second, the pool of items gathered in preparation for the study were the raw materials from which to construct the tests for each category of objectives. Third, the study gave us information about which categories of objectives were valued most highly and least highly by the English teachers, the actors, and the other groups.

[5] A full report of this study may be found in James Hoetker and Richard Robb, "Drama in the Secondary English Class: A Study of Objectives," *Research in the Teaching of English* (Fall 1969), pp. 127–159.

CHAPTER TWO

THE EXPERIMENTAL DESIGN

Trying to explain the design of this experiment to the lay reader, i.e., the reader who does not have at least a nodding acquaintance with the language of the scholarly specialty known as experimental design, is rather like trying to explain film speeds to someone who has never taken a photograph. Experienced methodologists have advised me that the effort can lead only to mutual frustration. The level of thinking about research within the educational community, they have told me, is so primitive that there is no point in even trying to talk to most educational researchers about experimental designs.

However, the attempt to explain the logic of the design must be made, for at present specialists in research methodology speak only amongst themselves. The majority of educational researchers continue to muddle along unaware even of the existence of experimental techniques which have for years been commonplace in such fields as agriculture, the biological sciences, and psychology. The fact that there is no communication between the methodological theorists and the working researchers has produced a situation in which much time and money is wasted on experimental studies which are of practical value primarily to aspiring methodologists who spend their time tearing inferior studies to pieces in journals.

Our concern here, however, is not with research studies which are simply faulty, for example, those which involve biased samples or inappropriate statistical analyses. Rather our concern is with studies which are representative of the best research that has been done in English, studies which are technically sound but methodologically inadequate. Rather than criticize the work of any individual, let us describe a typical study of the better sort and then discuss the ways in which it is less than adequate to its purposes.

Assume that we wish to evaluate a highly touted new technique for teaching written composition. We randomly divide our student subjects and our teachers into three groups: an experimental group which is to use the new method, a control group which is to use a conventional method, and a placebo group which is to do something unrelated to written composition. We

give all three groups a pretest. Then, after the experimental and control groups work for a time according to the prescribed methods, all three groups are given a posttest. Then the differences between the three groups are tested for significance, probably using analysis of variance or covariance.[1]

In what ways is this design inadequate? First of all, it is inadequate in its global conception of the experimental variables. A method of teaching written composition is a very complex phenomenon. One might identify any number of dimensions along which the experimental and the conventional methods differ from one another and as many dimensions along which they do not differ in any important way. Whether the results of the experiment are positive or negative,[2] we learn very little about what

[1] Several predictable patterns seem to govern the reporting of such studies. If the differences are in favor of the experimental group, the experimenter will, according to his temperament, make great things of it or cautiously suggest that, of course, further research is called for. If the results favor the control group two things may happen, depending upon the experimenter's personal commitment to the new method. If the experimenter is neutral, he will simply report that there is no evidence in favor of the new method. If he is deeply committed to the new method, chances are he will become the harshest critic of his own procedures and seek out reasons why his experiment did not demonstrate the superiority of the method that is self-evidently superior. If the control or the placebo group makes the highest score, then one of two things may happen. If the study is a short and inexpensive one, it will probably be filed away and forgotten. If the study involved a considerable investment of the experimenter's time, then there will be an intense effort to explain away the findings.

If the analyses of the data show that there is no difference between the methods—and for many reasons this is the result to be expected from any educational experiment—then the experimenter will be obliged to indulge in a ritual known as explaining negative findings. This involves identifying the many factors that might have masked real effects or produced spurious effects. The explanations are so familiar that they might economically be printed up in a standard chapter that could be appended without alteration to most reports of experiments or even more economically be referred to by a number or short title.

[2] Any experimental manipulation of programs, curricula, methods, or administrative procedures is almost certainly going to exert a weaker influence upon a student's performance at a particular time than that exercised by his entire previous life history, so the most sensible prediction for any experimental or evaluative study is *no difference*. Even if the design is sound, the measurements sensitive, and the experimental treatment pedagogically superior, the experimental treatment is, as J. M. Stephens puts it, "one slight change, imposed on a whole battery of powerful, prior forces," and it "may have great difficulty in demonstrating its influence." (J. M. Stephens, *The Process of Schooling: A Psychological Examination* [New York: Holt, Rinehart and Winston, Inc., 1967], p. 85.) In this book

parts of the treatments had what effects. Let us say, for example, that the experimental and conventional methods differ from one another in the following theoretically important ways:

Area of Difference	*Experimental*	*Conventional*
1. Classroom organization	Student-centered	Teacher-centered
2. Source of topics	Personal experience	Textbook
3. Primary writing activity	Creative writing	Essays on assigned topics
4. Frequency of writing	As students wish	Once a week

Now it may well be the case that only one of these differences has an important effect on written composition scores. If, for instance, classroom organization were so powerful an influence that the student-centered classes scored significantly better than the conventional classes, the experimenter would have no way of knowing that only the one element of the experimental treatment was in fact superior to its counterpart in the conventional treatment. He would be in great danger of building a spurious case for the overall superiority of the new method, perhaps emphasizing the importance of an element that was in fact not important.

To take another possibility, it is conceivable that classroom organization affected student writing ability in one direction while the frequency of writing affected it in the other. In this case two important influences might cancel each other out and the results of the experiment falsely suggest that the two methods were indistinguishable in their effects. The experimenter simply cannot tell what differences between treatments are the important or effective ones. So the first point to be made is that our knowledge is unlikely to be advanced by experimentation until such time as we utilize designs which enable us to

Stephens summarizes (pp. 71–92) the results of several thousand studies of classroom learning and concludes they show that students learn something no matter what the schools do and that none of the factors that have been studied have shown to affect student learning in any consistent way. The theory of "spontaneous schooling" which Stephens advances to help explain the negative results of research studies in education is provocative and should be familiar to anyone involved in program evaluation and educational research. Basically, Stephens argues that those things which are pedagogically most important—e.g., immediate reinforcement of a student response by an unconscious alteration in a teacher's expression—have not been, and perhaps cannot be, manipulated in experimental studies.

get beyond global definitions of our variables and enable us to examine the effects and interactions of the constituent elements of the treatments with which we are concerned.[3]

The second inadequacy of the typical experiment has to do with its lack of control of unmeasured variables that may influence the results. Random assignment of subjects to conditions only assures there will be no systematic biasing of the results. It does not really control for between-group differences that can at times be more powerful determinants of performance than the treatments being evaluated in the experiment. This is especially true of experiments in the schools, where the experimenter is rarely able to assign individual students to treatments but must work with intact groups that have been previously constituted by unknown administrative procedures.

The possibility of radical differences between randomly assigned groups is only one of a host of factors which cannot be

[3] Another way of making the same point would involve contrasting typical weak models, which deal with total variance estimates, with strong models, which enable the experimenter to partition the variance so that he can, in evaluating differences between levels of an independent variable, deal only with that portion of the variance attributable to the independent variable in question. The prediction of negative results for any experimental study (cited in the previous note) applies only to weak experimental models —those in which the total variance in performance scores is involved in the contrasts. It does not apply with equal force to stronger models. For example, if 95% of the performance differences between groups of students at two levels of an experimental variable are due to unmeasured random factors, then it is of course unlikely that the effects of any experimental factor will be great enough to produce significant differences between levels if a weak model is used, since the effects of the experimental factor are, as it were, lost in the noise made by the random factors. With a strong model, however, it is theoretically possible to partial out the 95% of the variance due to random factors and to deal only with the differences in student performance that are due to differences between the levels of the independent variable in question. In practice, of course, it is never possible to control for, or estimate, all random sources of variation.

But educational researchers must inevitably deal with weak factors and small effects, and they must get out of the habit of thinking in terms of crucial tests of competing hypotheses. Paradoxically, the weak typical experiment is appropriate only when theory, measurement, and techniques for manipulating the experimental variables are very far advanced, as in the physical sciences. In educational experiments strong models are essential so that (1) real and possibly important effects can be detected, (2) no-difference conclusions will not be reached when there are indeed differences, and (3) no-difference findings may be taken as dependable evidence that the effects of different levels of the independent variables are indistinguishable.

taken account of in the conventional experimental-control, pretest-posttest type of design. Analysis of covariance procedures can at best control for only a few extraneous factors. So no matter what the results of such an experiment, there will remain any number of plausible alternative explanations for the results, alternative explanations which the design can do nothing to rule out or control for. Speculations about alternative explanations are the stuff from which final chapters and critical reviews are made. But this type of post facto speculation is of little value to anyone. What is needed is for speculation about alternative explanations to take place before the designing of a study is undertaken. Our knowledge is unlikely to be advanced by experimentation until such time as we take into account in our experimental designs precisely those factors that we have traditionally relegated to speculative discussions of negative results and critical reviews of published studies.

The present experiment, the design of which will be discussed below, goes a long way toward overcoming both these major inadequacies of the typical study. It simultaneously evaluates the effects of a number of factors which are elements in the treatments being compared, and it controls for the influence of most of those factors, aside from the ones being evaluated, which might affect student performance.[4]

These differences being crucial, it seems important to try to explain in detail how this experiment differs from the typical experiment described above. The discussion below is as free from jargon as possible, but there are certain unfamiliar terms which cannot be dispensed with.

We will assume a reader familiar with basic statistics and the standard literature on research, but we will, at the risk of seeming to patronize, start out by defining some basic terms. A variable is anything which exists in more than a single state,

[4] In the typical experiment, unless serious procedural errors have been made, one may have some confidence in his positive findings, if only on the grounds that a factor must be powerful in its influence if it can overcome the multitude of other factors working toward a no-difference finding. But in the typical experiment negative results are not very informative, since they may mean only that the treatment effects were overshadowed by the effects of unmeasured factors. When, however, the extraneous factors are accounted for, as in the present design, negative results are informative, and it is possible to interpret a no-difference finding with some confidence as meaning that a factor did not have a significantly large effect.

anything which can vary. An independent variable or treatment variable or factor is one which is manipulated in an experiment, e.g., a teaching method used with one of several groups in a comparative study. A dependent variable or criterion measure or dependent measure is a variable which varies presumably as a function of changes in an independent variable. In the typical experiment described above, the dependent variable was a test score of some sort used to measure the effects of three variations in teaching method.

These variations were, you will recall, an experimental method of teaching composition, a conventional method of teaching composition, and the study of something unrelated. Most researchers would refer to the experiment as involving a comparison of the effects of three independent variables. But—and this is crucial to an understanding of everything that follows—it is more useful to conceive of the experiment as evaluating the effects of three levels of a single independent variable called teaching method.

A variable may be spoken of as having any number of levels. The division of a variable into levels may be naturalistic (before-after; night-day) or arbitrary (high I.Q.-low I.Q.; high, low, moderate manifest anxiety). In the case of the present study, the variable of timing, as discussed in the previous chapter, has two levels: study before the performance and study after the performance. In an experiment which was concerned only with the effects of timing upon the test scores being used as a dependent measure, we would have an experimental design which incorporated two levels of an independent variable called timing. It is conventional when an independent variable has two levels to refer to one level with a plus sign (+) and the other level with a minus sign (−). In planning the analyses we would speak of contrasting the scores of subjects at the + level with the scores of subjects at the − level.[5]

[5] Actually this is not a notational convention, but a system of weighting scores at different levels of a factor. With a two-level factor, the weights +1 and −1 may be assigned to the levels; with a three-level factor, the weights might be +1, 0, and −1, and so on. Say the mean scores on a test used as a dependent variable were 45.5 and 51.0 for the two levels of a particular factor. If the levels were weighted +1 and −1, respectively, the sum of the weighted mean scores would be $+1(45.5)-1(51.0) = 5.5$, and the question would be whether, in the particular circumstances, 5.5 is signifi-

However, the present experiment was not concerned with levels of a single variable, but with the various combinations of the levels of several variables. Let us introduce a second variable of content of the lessons; it also has two levels which we can call specific to the text and related to the text. We wish to consider in a single experiment both the timing and the content of the lessons used in conjunction with a performance of a play. What will be manipulated in the designing of the experiment are the levels of these two variables. With two two-level variables there are $2^2 = 4$ possible combinations of levels as follows:

Table 1

A 2^2 Factorial Design

Run	*Timing*	*Content*		*Run*	*Timing*	*Content*
1.	+	+		1.	Before	Specific
2.	−	+	OR	2.	After	Specific
3.	+	−		3.	Before	Related
4.	−	−		4.	After	Related

The sort of experimental design we now have is called a factorial experiment, which simply means an experiment in which two or more treatment variables are evaluated simultaneously.[6] The particular factorial experiment above would enable us to look at the effects of the interactions between the two treatment variables in question, already a considerable advance over the typical design, since in education it is very likely that no independent variable is so powerful in its effects that it will not be influenced by other variables.

Let us take this one step further and introduce the variable,

cantly different from zero. For the purposes of this presentation, however, the + and − signs may be considered simply as a shorthand way of distinguishing one level of a factor from the other.

[6] A good brief introduction to the logic of factorial designs is in Frederick N. Kerlinger, *Foundations of Behavioral Research: Educational, Psychological, and Sociological Inquiry* (New York: Holt, Rinehart and Winston, 1964), pp. 322–336. There are any number of excellent textbook treatments of the subject available to anyone with a knowledge of basic statistics. Roger E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences* (California: Brooks/Cole, 1968) is probably the best, however, for someone trying to instruct himself. Chapters 11, 12, and 13 in Allen L. Edwards, *Experimental Design in Psychological Research* (New York: Holt, Rinehart and Winston, 1968) are also extremely useful.

intensity of treatment of the text. If we arbitrarily define two levels of this variable as brief and intense, we may then design an experiment evaluating all combinations of the levels of the three variables. A three-variable factorial experiment in which all the variables have two levels has $2^3 = 2 \times 2 \times 2 = 8$ possible combinations of the levels of the variables. The experimental design itself would be referred to as a 2^3 factorial experiment, and an evaluation of all the combinations would require eight runs, or subjects. Using the + and − symbols, all the combinations of levels of the three major independent variables in the 2^3 factorial experiment would be represented by the following design matrix:

Table 2

A 2^3 Factorial Design Matrix

	Timing	*Content*	*Intensity*
1.	+	+	+
2.	−	+	+
3.	+	−	+
4.	−	−	+
5.	+	+	−
6.	−	+	−
7.	+	−	−
8.	−	−	−

At this stage of designing the experiment all we are talking about is describing the run or the treatment condition for each group of subjects in terms of particular combinations of levels of the independent variables that we are interested in. I hope that by this stage the principle is clear: when, as in most realistic cases, more than one two-level independent variable is of interest, all possible combinations of the levels of the independent variables can be evaluated in a number of runs equal to two raised to the power of the number of variables.

Let us go a step further then. In the experiment being reported here there were actually five independent variables of interest which we wished to evaluate simultaneously. The variables and the signs given to the two levels of each are summarized in Table 3.

With five two-level variables, there are 2^5 possible combinations of levels and it will require thirty-two different groups of

Table 3

Summary of Variables and Levels of Variables in the Experimental Study

Variable Name	*Levels*	*Sign*
A. Background study	Brief	−
	Intensive	+
B. Textual study	Brief	−
	Intensive	+
C. Timing of lessons	Before performance	+
	After performance	−
D. Content of lessons	Related to play	−
	Specific to play	+
E. Play performance	Attend	+
	Not attend	−

subjects to try out all the variations. If it is desirable, and it usually is, to have two or more subjects or groups of subjects in each of the runs, then it would require a minimum of sixty-four subjects to obtain all the desired estimates. But sometimes it is possible to reduce the number of subjects required without losing any information of interest. This may be done by using only one level of one or more of the independent variables. A design which uses only a fraction of the possible combinations of the levels of the variables is known as a fractional factorial design.[7]

In the present case we had no immediate interest in the level of the play performance variable called not attend (E in Table 3). The hypotheses in dispute between the actors and educators had to do with the interactions of classroom treatments

[7] The standard treatment of fractional factorial designs is the monograph by G. E. P. Box and J. S. Hunter, *The 2^{k-p} Fractional Factorial Designs* (University of Wisconsin, Mathematics Research Center, United States Army, Technical Summary Report #218, 1961). Chapter 10 in Kirk's *Experimental Design* is also excellent, although his system of notation is less elegant than Box and Hunter's. Kirk gives a list of references to studies that have used fractional factorial designs. The National Bureau of Standards of the U. S. Department of Commerce has published in its Applied Mathematics Series pamphlets in which are summarized all varieties of fractional factorial designs at two and three levels. The pamphlet numbers are 48 and 54, respectively, and they are available from the U. S. Government Printing Office.

with attendance at a performance of a play. Consideration of the classroom treatments apart from the performances could wait. We could therefore use only one + level of the play performance variable. Using only the + level of the play performance variable in the design reduces the number of runs necessary to $2^{5-1} = 2^4 = 16$. The matrix describing the resulting design is given below. Technically it would be called a one-half replication of a 2^5 factorial experiment. The missing half of the design would be a duplicate of the one in Table 4, but with 16 minus signs in column E.[8]

Table 4

Design Matrix for a 2^{5-1} Fractional Factorial Design

Run Number	*Variable*					*Run Number*	*Variable*				
	A	*B*	*C*	*D*	*E*		*A*	*B*	*C*	*D*	*E*
1	−	−	+	−	+	9	−	+	+	−	+
2	+	+	−	+	+	10	+	−	−	+	+
3	−	−	+	+	+	11	−	+	+	+	+
4	+	+	−	−	+	12	+	−	−	−	+
5	−	−	−	−	+	13	−	+	−	−	+
6	+	+	+	+	+	14	+	−	+	+	+
7	−	−	−	+	+	15	−	+	−	+	+
8	+	+	+	−	+	16	+	−	+	−	+

By referring to the summary in Table 3 it is possible to read off from this matrix a description of the experimental treatment that will be given to the subjects in each run. For example, the classes in run number one will have a brief study of the background (− level of A) and a brief study of the text (− level of B) before attending the performance (+ level of C), and the content of the lessons will be related to the play being performed (− level of D). Though it is confusing at first, the

[8] It would obviously have been possible, and simpler, to explain the design as a four-factor full factorial experiment, rather than as a 2^{5-1} fractional factorial. Formally, the procedures for analyzing the data from a 2^{5-1} design are identical to those for analyzing data from a 2^4 design. But the interpretation of the results in the two cases is quite different. The consequences of conceiving of the design as a 2^{5-1} experiment are explained in Chapter Five. For the moment suffice it to say that from the first, the researchers working on this study thought of it as an experiment involving five factors, one of which was attendance, or non-attendance, at a play, so the treatment of the design in this chapter is simply historically accurate.

utility of this system of notation will be obvious if one only tries to think or write about a factorial design without resorting to some such shorthand.

At this stage we have an experimental design which enables us to evaluate not only the effects of different levels of each of the independent variables, but also to evaluate any number of interactions between the levels of the variables. But the design described by the foregoing matrix is still open to the objection that scores on the dependent measures are going to be affected in unknown ways by a host of unmeasured variables: class I.Q., prior theatre experience, the social structure of the classroom group, teacher rapport with the students, teacher knowledge of theatre, social and ethnic homogeneity of the class, etc. So it is desirable that we control for these factors or find a way to estimate their effects.

There are several general strategies, supplemental to random assignment of classes to treatments, for taking account of such factors. The first would be to devise measures for those variables considered likely to be important and introduce these variables as independent variables in the design. For example, one could get I.Q. scores and prior-theatre-experience scores from each class involved in the experiment, reduce these scores to two-level (high-low) variables, and incorporate them in the experimental design as the sixth and seventh independent variables. But this would yield a $2^{7-1} = 64$ run design and still leave unaccounted-for all the other possibly important unmeasured factors.

A second strategy would be to get measures on the potentially important variables and to statistically control for their influence. We used this strategy in regard to verbal intelligence and prior theatre experience because we had reason to believe that those two factors would be most likely to interact with the treatment variables. By using this strategy we denied ourselves the chance to examine interactions between I.Q. and the other variables. And we were still left with the possibility open that a part of the variance in scores on the dependent measures would be attributable to variables other than those included in the experiment.[9]

[9] Another strategy would involve using scores on the most important factors to assign subjects to blocks. This tactic was not available to us in

A third strategy available would involve repeating the entire experiment in such a way that the effects of the unmeasured variables would be indistinguishable from other effects. This can be best understood by referring to Table 5, which is a representation of the final complete design for the study. The whole experimental design has two blocks, the first an execution of the 2^{5-1} design in connection with the first play presented by the Project and the second a repetition of the design in connection with the second play. The two blocks are identical, except that the numbers in the right-hand column have been folded over; each group of subjects is assigned to a second block treatment that is the mirror image of its first block treatment. In the first block of the experiment, for example, subjects identified as 8 engage in intensive study of both the text and background of play-related materials before they attend the performance. In the second block of the experiment the same subjects engage in brief study of both the text and background of the play itself after they have seen the performance. What this means is that each group of subjects is contrasted with itself.

A simple example may make clearer the principles involved in the design. Imagine you are a contractor who needs to purchase a number of hammers. Two types of hammers are available, and the maker of each claims that his design enables a workman to drive more nails per minute. You wish to put the claims to an experimental test. In the terms we have been using, the independent variable is type of hammer and its two levels are Essex hammer and Bangrite hammer. You find two carpenters, give each one of the experimental hammers, and ask them to drive as many nails as they can in one minute. The dependent measure is the number of nails driven. Let us say you get these results:

Carpenter	*Type of Hammer*	*Number of Nails*
Bill	Essex	32
John	Bangrite	20

regard to the intelligence factor, since most available classes were not tracked by ability and there was not enough time between the opening of school and the start of the experiment to administer I.Q. tests and then choose classes of subjects on the basis of the results of those tests. The same considerations would have prevented us from using I.Q. as an independent variable even if we had wished to.

Table 5

The Design for the Experimental Teaching Study (First Version)

BLOCK 1 = FIRST PLAY				
Timing	*Content of Lessons*	*Intensity*		*Subject ID Number*
		Background	*Text*	
Before attending performance	Play-related	Intense	Intense	8
		Intense	Brief	16
		Brief	Intense	9
		Brief	Brief	1
	Play-specific	Intense	Intense	6
		Intense	Brief	14
		Brief	Intense	11
		Brief	Brief	3
After attending performance	Play-related	Intense	Intense	4
		Intense	Brief	12
		Brief	Intense	13
		Brief	Brief	5
	Play-specific	Intense	Intense	2
		Intense	Brief	10
		Brief	Intense	15
		Brief	Brief	7

BLOCK 2 = SECOND PLAY				
Timing	*Content of Lessons*	*Intensity*		*Subject ID Number*
		Background	*Text*	
Before attending performance	Play-related	Intense	Intense	7
		Intense	Brief	15
		Brief	Intense	10
		Brief	Brief	2
	Play-specific	Intense	Intense	5
		Intense	Brief	13
		Brief	Intense	12
		Brief	Brief	4
After attending performance	Play-related	Intense	Intense	3
		Intense	Brief	11
		Brief	Intense	14
		Brief	Brief	6
	Play-specific	Intense	Intense	1
		Intense	Brief	9
		Brief	Intense	16
		Brief	Brief	8

After this has been done, however, there remains the possibility that this difference does not mean the Essex hammer is superior, but that the workman using it is stronger or more skillful. Let the carpenters exchange tools and repeat the experiment. Following is one possible outcome of the two replications:

Carpenter	*Order*	*Type of Hammer*	*Number of Nails Example A*
Bill	First	Essex	32
	Second	Bangrite	24
John	Second	Essex	28
	First	Bangrite	20

The total nails-per-minute score for the Essex hammer is the sum of Bill's thirty-two nails plus John's twenty-eight nails; the total for the Bangrite hammer is the sum of Bill's twenty-four nails plus John's twenty nails. Note that the same subjects contribute scores to the total score associated with each level of the independent variable. The fact that Bill seems to be about four nails per minute faster than John regardless of the tool being used does not significantly affect the contrast. Which is to say that in this particular case the Essex hammer seems to be the superior design no matter which workman is using it.

Two more of the possible outcomes of such an experiment are these:

Carpenter	*Order*	*Type of Hammer*	*Number of Nails Example B*	*Example C*
Bill	First	Essex	32	32
	Second	Bangrite	32	20
John	Second	Essex	20	20
	First	Bangrite	20	32

According to the figures in Example B, the nails per minute rate for the four runs are:

	Type of Hammer	
Carpenter	*Essex*	*Bangrite*
Bill	32	32
John	20	20

The mean nails-per-minute rate is the same for each level of the independent variable called type of hammer; all the variation between cells seems to be due to the fact that Bill can for some reason drive nails faster than John. It is a characteristic of factorial designs that they enable one to look at the main effects of the independent variables, e.g., carpenter or type of hammer, separately and to look at the interactions between the variables as well. Example C illustrates what is meant by interaction between the independent variables.

	Type of Hammer	
Carpenter	*Essex*	*Bangrite*
Bill	32	20
John	20	32

The mean nails-per-minute rate for each type of hammer is the same, but the table shows that Bill is superior to John while using the Essex hammer and that John is superior to Bill while using the Bangrite hammer. Here we have an interaction between the workman and his tools.

Compare the knowledge gained in these three cases with that gained from the one-shot comparison between hammers. From the typical experiment one would conclude that the Essex hammer was superior and would presumably order a batch of them. The experiment in Example A would, as it happens, confirm the superiority of the Essex hammer, but would give us more faith in the result and a better idea of the true difference in nails-per-minute rates of the two hammers. The experiment in Example B would lead us to conclude that the differences in nails-per-minute rates were due entirely to the skills of the carpenters and that we would have to do more research before we could decide which hammer to buy. The experiment in Example C suggests that there is a difference between hammers, but that we will want to order Essex hammers for workmen like Bill and Bangrite hammers for workmen like John.

This example may also be used to preview two important points that will be fully discussed later. First, the original experiment contrasted the nails driven by Bill using the Essex hammer with the nails driven by John using the Bangrite hammer. In this case we could say that the carpenter effects and the hammer effects were confounded, which is to say that they

are inseparable or indistinguishable. It has been shown that one of the primary advantages of a factorial design is that it enables us to evaluate these effects separately and in interaction with one another. But when one uses a fractional factorial design, he loses part of this advantage, as he must confound certain effects with others and thereby lose some information.

Second, although a single dependent variable was used in the example, any number of dependent variables may be used in such an experiment. Our hypothetical contractor might have wished to measure the number of strokes per nail, amount of noise made, the drops of perspiration on the carpenters' foreheads, the number of hits upon thumbnails, and the obscenity-per-minute rates. In the study that we conducted there were thirteen dependent variables measuring different aspects of the response to drama.

It is very likely that the scores on several of these dependent variables will be correlated with one another. The carpenter who hits his thumb the most often, for example, is also likely to drive the fewest nails, produce the most perspiration, and have the highest profanity rate. If, using the procedures familiar to most researchers, the effects of the independent variables are evaluated for each of the dependent variables separately, the analyses may give a series of misleading significant differences.

Say that the Bangrite hammer is badly designed, so that one using it will hit his thumb significantly more often than one using the Essex hammer. Thumb hitting will be negatively correlated with the nails-per-minute rate and positively correlated with perspiration and cursing. Let us assume that when the between-hammer differences are evaluated they are highly significant on all four of these variables.

Now the chances are that the perspiration and cursing rates are not in any important sense related to differences between hammers. But from a conventional, or univariate, analysis, one cannot tell whether this is the case or not. One is in danger of concluding that buying Essex hammers will reduce the number of passersby who will be shocked by the carpenters' language or their aroma, when the more appropriate interpretation would be that amount of perspiration and choices of expletives are matters of individual differences, related to the number of in-

Table 6

The Design Matrix for the Experimental Teaching Study
(Second Version)

Run	*Order*	*Background*	*Text*	*Timing*	*Content*
1	First Play	−	−	+	−
	Second Play	+	+	−	+
2	Second Play	−	−	+	−
	First Play	+	+	−	+
3	First Play	−	−	+	+
	Second Play	+	+	−	−
4	Second Play	−	−	+	+
	First Play	+	+	−	−
5	First Play	−	−	−	−
	Second Play	+	+	+	+
6	Second Play	−	−	−	−
	First Play	+	+	+	+
7	First Play	−	−	−	+
	Second Play	+	+	+	−
8	Second Play	−	−	−	+
	First Play	+	+	+	−
9	First Play	−	+	+	−
	Second Play	+	−	−	+
10	Second Play	−	+	+	−
	First Play	+	−	−	+
11	First Play	−	+	+	+
	Second Play	+	−	−	−
12	Second Play	−	+	+	+
	First Play	+	−	−	−
13	First Play	−	+	−	−
	Second Play	+	−	+	+
14	Second Play	−	+	−	−
	First Play	+	−	+	+
15	First Play	−	+	−	+
	Second Play	+	−	+	−
16	Second Play	−	+	−	+
	First Play	+	−	+	−

juries sustained but not directly related to the type of hammer being used.

Multivariate analysis of various procedures such as those used in this study enable one to evaluate all of the dependent variables simultaneously, so that the former sorts of conclusions may be avoided and the latter sort reached. To illustrate, in a multivariate analysis the following questions about the effects of the type of hammer would be asked in turn: (1) Is the number of nails driven significantly different between the two types of hammers? (2) After the variance due to number of nails driven has been taken out, is the variance in number of hits on the thumb different between hammers? (3) After the variance due both to number of nails driven and hits on the thumb has been taken out, is there a difference between hammers in the rate of profanity emission? (4) After variance due to the three preceding variables has been taken out, is there a difference between hammers in the amount of perspiration produced? Proceeding in this way one may distinguish differences due to variations in the independent variable from those due to intercorrelations between the dependent variables.

There will be further discussion later of the multivariate analyses used in this study, but this illustration may serve to orient the reader and to put the discussions of the experimental design and of the dependent variables into a more meaningful context.

To return to our own design, Table 6 presents the design matrix for the entire experimental study in symbolic form. The contrasts of primary interest will be those involving the total scores on each dependent variable, i.e., the first block score plus the second block score. But it will also be profitable to examine additional contrasts, especially those within and between blocks and within categories of tests and to combine the dependent variables in a number of ways.

CHAPTER THREE

DEFINING THE INDEPENDENT VARIABLES

After the variables to be involved in the study had been identified and the design completed, members of the CEMREL staff went to Providence, Rhode Island, for a two-day meeting in June 1968, with approximately fifty tenth-grade English teachers from all over the state. The meeting was also attended by administrative personnel of the Educational Laboratory Theatre Project, representing both the Trinity Square Repertory Company and the schools, and by representatives of the Rhode Island State Department of Education.

The purpose of the experimental study was explained, and the experimental design was presented and discussed in general terms. Categories of dependent variables were suggested on the basis of the first analysis of the data from our study of objectives for drama. The teachers then were asked to make two contributions to the planning of the study. The first was to define the independent variables in terms that were realistic and meaningful to them as English teachers. The second was to contribute items which might be used on tests constructed to measure each of the dependent variables we had identified.[1]

[1] We consider involving the teachers at this stage of the planning of the experiment to be of the utmost importance. The operationalizing of the experimental variables is the responsibility of the practitioners and subject matter specialists, and their needs and their judgments must sometimes take precedence over the preferences of both the methodologist and the psychometrician; for it is when the variables are operationalized by scientists untrained in the discipline being studied that the experiment is likely to be concerned with trivialities or unrealistic and uninteresting contrasts.

The involvement of the teachers not only gave us definitions of the variables that were sensible and significant to working English teachers, but also gave the teachers a stake in the experiment. Furthermore, since each of the teachers who was to help carry out the experiment had had a voice in planning it and since each of them understood that each of the treatments had to be carried out in a particular way if the experimental results were to be interpretable, the teachers were willing to abide by the specifications of the treatment conditions even when, as was often the case, a particular treatment went against a teacher's best judgment about what should be done. The importance of this cannot be overemphasized, since two of the things which traditionally have plagued methods experiments covering long periods of time have been attrition, resulting in an uninterpretable biasing of the experiment, and the departure of experimental teachers from the procedures that the experiment is supposed to be evaluating.

At the meeting the consensus was quickly reached that the questions to be investigated in the proposed study were both crucial to the Project and important to English teachers; that the variables in the proposed design were indeed the important ones; and that it made sense to consider each of the variables as dichotomous or two-leveled. Each of the independent variables was discussed in turn, and by the end of the second day each of the levels of the experimental variables had been described in concrete terms to the satisfaction of the teachers, the Project officials, and the experimenters. The definitions that were arrived at are described below.

Timing

The two levels of the timing variable were before the performance and after the performance. But the further specification was made that before treatments should be scheduled so that they would be completed on the school day before students attended the theatre, while after treatments were to begin on the day succeeding the performance but following a period of time allowed for free discussion of the play.

Plays

At the time of this first meeting the titles of the first two plays that would be presented during the following season were not known. It was certain only that the second play would be one by Shakespeare. However, it was possible to decide that the treatment variable to be called play attendance should for the sake of uniformity be considered as consisting of theatre attendance plus approximately one-half hour during the immediately succeeding class period which was to be devoted to spontaneous reactions to the performance. In other words, this discussion period would be, like the play itself, common to all treatment conditions. It was thought wise to make this stipulation since it was often difficult to keep students from talking about the plays, and if some teachers prevented such discussion while

Only one teacher was lost in the six-month course of the present experiment.

Items which asked students to report on the length and content of the lessons and the methods used by the teacher revealed almost no variation between what the teachers had agreed to do and what their students reported them doing. The involvement of the teachers in the planning does not by itself account for this remarkable set of circumstances, but we think it did contribute importantly to the quality of the study.

others allowed it two treatment conditions which were the same on paper might be different in fact.

Content

It was first agreed that the play-specific level of the content variable should be defined in terms of materials included in the portfolios that were given to all English teachers prior to the performance of each play. The portfolios for the next season's plays were not yet available, but the Project administrators were able to assure the teachers that the portfolios would include a collection of biographical and background materials, notes by the director and other theatre personnel, a suggested study plan, and various other supplementary materials. It was also agreed that a copy of the play to be performed would be supplied to each student in a class at the play-specific level of the content variable.

It was further agreed that the play-related level of the variable would be defined in terms of the experimental *An Introduction to Theatre* lessons which had been developed at CEMREL in connection with the Project.[2] A good number of the teachers present at the meeting had used or were familiar with these materials and some had helped to plan them. It was desirable to have a set of standard materials at the play-related level, so that the levels of the background and text variables could be defined in terms of materials from those lessons and from the portfolios. CEMREL agreed to supply teachers and students at related levels with all necessary materials and books. But as was brought up at the meeting the use of the CEMREL drama lessons would produce some confusion. The drama lessons, two volumes of which were available at this time, had been designed to help English teachers approach drama through the medium of dramatic activities and to introduce a new dimension into the classroom study of drama. Therefore the use of these materials

[2] These curriculum materials were developed specifically for the Project in the attempt to devise a method for assisting English teachers untrained in drama to deal with the theatrical aspects of the plays being presented in the Project. The general title of the series of lessons is *An Introduction to Theatre*. Two volumes of the lessons were available at the time of the experiment: James Hoetker and Alan Engelsman, *Reading a Play* (St. Louis: CEMREL, Inc., 1968) and James Hoetker, *Shakespeare's* Julius Caesar: *The Initial Classroom Presentation* (St. Louis: CEMREL, Inc., 1968).

would confound the effects of studying related materials with the effects of teaching drama through dramatic activities. A parallel confounding at the play-specific level of the variable could be introduced, however, by specifying that the play-specific level would not involve dramatic activities, but would deal with the text of the play in the analytical manner conventional in most English classes.

The consensus of the teachers was that the advantages of having standard materials outweighed the difficulties of interpretation introduced by the confounding of materials and methods. That is to say, the contrast between the play-specific and play-related levels would still involve classes which had studied the play and classes which had not studied it. If it should happen that the play-related conditions produced higher scores on a number of dependent measures, then it would be time to design another experiment in which the materials and methods were studied separately. This study then is not directly a test of the CEMREL drama curriculum or a comparison between dramatic and analytical methods of studying plays. (In certain cases, however, the experimental results enable us to make some suggestions about how methods and materials might have operated to give the observed results.)

Background and Text

It was decided that the levels of the background and text variables should be defined in terms both of the amount of material covered and the amount of class time expended. It was necessary in defining these variables to consider the levels in connection with the levels of content.

Intensive-Specific. Using all or most of the background material that is included in the portfolio, the students at this level are to spend from four to seven class periods studying the background of the play. The specified time includes time spent on library and research assignments.

Brief-Specific. Using one or two items of background material from the portfolio, students at this level will spend less than two periods studying the background of the play and will do no out-of-class research work or reading. The particular items to be used were to be specified by the Project Coordinator when the portfolios were completed.

Intensive-Related. Using the first volume of CEMREL's drama lessons in connection with the first play and the second volume of lessons with the second play, students at this level will spend from four to seven days studying backgrounds. In the first case this background would be a general orientation to the theatre; in the second it would be an introduction to Shakespeare by way of working dramatically with key scenes from *Julius Caesar*. *Julius Caesar*, by the way, had been presented the previous season, so we knew that our play-related conditions would not be transformed into play-specific conditions.

Brief-Related. Using particular lessons chosen by the authors of the CEMREL materials, students at this level will spend less than two days on an orientation to theatre in connection with the first play or on Shakespeare with the second play.

The operationalization of the levels of the text variable followed the same logic used to define the levels of the background variable. An intensive study covered four to seven periods; a brief study covered less than two periods. In the play-specific condition the intensive level read plays that were being performed; the first was Sean O'Casey's *Red Roses for Me* and the second was *Macbeth*. In the brief-specific condition the students read and discussed a single scene from the play in question. The related treatments for the O'Casey play were these. Students at the intensive level read and acted portions of Sean O'Casey's *The Plough and the Stars*. The students at the brief level worked dramatically with a cutting from *The Plough and the Stars*. The related conditions for *Macbeth* involved students at the intensive level in working dramatically with *Julius Caesar*. Those at the brief level worked with a single scene from *Julius Caesar*.

The portfolios for each play were prepared some weeks before each play opened for students. When they were ready it was possible to define each treatment condition very precisely. Before the first play each teacher participating in the experiment was randomly assigned to a treatment condition and given a package containing a sheet describing the experimental procedures he was to follow with the class he had chosen to participate in the experiment along with the necessary teaching materials and tests. A similar sheet accompanied the materials provided prior to the second play. A sample copy of one of these sheets is included in Appendix C.

CHAPTER FOUR

THE DEPENDENT VARIABLES

It was clear from the start that a large number of dependent variables would enter into this study if it were going to speak to the hypotheses it set out to investigate. The reason that the different groups involved in the Project had different ideas about what should be done in classrooms was primarily that they valued differentially the objectives that such a project might be expected to achieve. That is, an actor and a teacher might agree that method one would give the highest scores on dependent variable X; but the actor might nevertheless advocate method two because he thought it would raise scores on dependent variable Y, which he considered much more important than X. Our study of objectives showed that English teachers valued most highly objectives involving what might be called philosophical insights and those involving knowledge of dramatic literature. They therefore tended to advocate the combination of treatment variables they had reason to believe would lead to student achievement in those areas. Actors valued most highly objectives having to do with maximizing the affective response to the performance itself and those having to do with the transformation of this excitement into appreciation for the arts. They therefore advocated the methods they saw as doing as little as possible to hinder the spontaneous communication between the acting company and the audience.

Ideally the selection of dependent variables in a study such as this would enable the experimenter to state at the end that treatment variation one gave the best results on the objectives valued by English teachers, variation two gave the results most valued by actors, and so on. What we have been able to do is not quite so neat, but, as will be shown, some of our results may be interpreted in such a form.

The Objectives-for-Drama Categories

When we came to the meeting with the teachers in June 1968, we had the preliminary analyses of the data from our study of the objectives various groups held for the teaching of drama.

The analyses suggested that the objectives fell into six important groups:

1. Affective response to the production
2. Knowledge of the play being performed
3. Development of critical and interpretive skills
4. Acquisition of philosophical and moral insights
5. Appreciation of literature, drama, and the arts
6. Development of desirable attitudes and behaviors

We discussed the study and this categorization with the teachers, and there was general agreement that the categories probably included most of the educational objectives that would be of interest to educators and theatre people. But a number of subcategories and subsidiary categories were suggested, and it became clear that the number of dependent measures was such that we were going to be restricted largely to the use of teacher-administered paper-and-pencil tests.

We asked the teachers at the meeting to take an hour to write items that might be used to test achievement in categories 1, 4, 5, and 6. Categories 2 and 3 would consist of items specific to the as-yet-unchosen plays. The items contributed by the teachers were added to the pool of several hundred items already collected in the course of preparing the study of objectives for drama. As might be expected, there was a great deal of duplication between the teacher-written items and the ones we had gathered from printed sources and written ourselves.

The task of constructing instruments to obtain measurements in each of the categories was begun immediately after the meeting with the teachers. Five members of the research staff spent several days working together, simultaneously considering the assignment of items to categories and methods of converting the items into easily administered tests. In the course of these deliberations several refinements were made in the categories. For example, the appreciation category was divided into subcategories called attitudes, cognitions, and discriminations on the basis of the content of the items originally assigned to that category. Other categories were divided on the basis that the several types of items in the category called for different types of student responses, so that, in effect, more than one test was constructed for a single dependent variable; two knowledge tests

were written, for instance, one involving true-false items and the other the identification of quotations. When the categories were set, a table of random numbers was used to select the items from each pool which would appear on a test. Writing and revising the tests themselves took several weeks more.

Discussion of the Dependent Variables

A total of fifteen dependent measures plus a number of other questionnaire-type items that were external to the design itself were finally used. Table 7 summarizes the titles of the dependent variables and gives the abbreviation of each title which was used for coding purposes and which will be used later in this report to conserve space. The X and Y prefixes indicate administration of the test in connection with the first play and second play, respectively. The abbreviation used without a pre-

Table 7

Titles and Code Designations of All Dependent Variables

Category	*Title*	*Code Designations*
1. Affective response	Liking for performance	XLIK, YLIK
	Involvement	XINV, YINV
2. Knowledge of play	Quotation identification *	XNOQ, YNOQ
	Factual knowledge (true-false) *	XNOT, YNOT
3. Interpretive skills	Interpretation	XINT, YINT
	Judgment of quality	XJUD, YJUD
4. Philosophical insights	Thematic understanding *	XPHI, YPHI
5. Appreciation	Attitudes	XAPA, YAPA
	Cognitions	XAPC, YAPC
	Discrimination	XADP, YADP
6. Desirable attitudes and behaviors	Attitudes	XDAT, YDAT
	Behaviors	XBEH, YBEH
	Theatre etiquette	XETQ, YETQ
7. Covariates	Verbal intelligence	VIQS
	Prior theatre experience	PREX

fix refers to the variable considered as the total score on the two administrations of the test, e.g., XLIK + YLIK = LIK. Those titles marked with asterisks designate tests made up of play-specific items, i.e., the X form of the test deals with *Red Roses for Me* and the Y form deals with *Macbeth*. In all other cases the X forms and Y forms of a particular test were identical. The tests described by these titles will be discussed below. One sample item from each test will be given to illustrate the form it took on the test.[1]

The Affective-Response Category

The first test in this category, liking for performance, consisted of a single question:

> Which of the following words or phrases comes closest to describing your own evaluation of the play that you just saw?
>
> A. Excellent
> B. Pretty good
> C. Uneven, sometimes good and sometimes poor
> D. Poor
> E. Very poor

Scoring was on the basis of one point for very poor through five points for excellent.

The involvement test consisted of thirty statements having to do with affective responses to a play in performance. Each student was to respond with an expression of how strongly he agreed or disagreed with the statement. There was no provision for a no-opinion answer, as shown in this example:

> I sometimes feel my heart beating faster when a play gets exciting.
>
> A. Strongly agree
> B. Agree
> C. Disagree
> D. Strongly disagree

Scoring was on the basis of one point for strongly disagree through four points for strongly agree for the positive items,

[1] Only sample items are included in the report since inclusion of all the tests would more than double the size of the paper. There were ten forms of each of the instruments: the pretests had six or seven pages, the postlesson tests had three pages, and the postperformance tests had six pages—a total of approximately 165 pages of tests.

and the opposite for negative items. There were twenty positive and ten negative items. The possible range of scores on this test was from 30 to 120.

The Knowledge-of-Play Category

The first of the two tests under this category involved quotations. There were twenty items, ten involving the identification of the speaker of the quotation and ten involving the identification of the character to whom the quotation was directed. The quotations chosen were, in our judgment, crucial or typical ones. The following example is from *Red Roses for Me:*

"Haven't you heard, old man, that God is dead?"

A. Brennan, the landlord
B. Mullcanny
C. Roory O'Balacaun

A correct identification was worth two points, so scores could range between 0 and 40.

The second test was a very conventional forty-item true-false test about the play: plot, characters, events, the facts. With one point for each right answer, scores could range from 0 to 40. An example follows:

Mrs. Breydon objects to Ayamonn's courting Sheila because Sheila is Catholic.

A. True
B. False

The Interpretive-Skills Category

The interpretation test consisted of ten anonymous quotations from prose, verse, and dramatic works. Two questions accompanied each quotation, and each question had five possible answers from among which the student was to choose the best. For example, the text of Emily Dickinson's "Much madness is divinest sense" was followed by these two questions:

The person speaking in this poem looks on madness as

A. Something only God can make sense of
B. A dangerous thing
C. A good thing

D. A bewildering condition
E. A form of insanity

The person speaking in this poem is probably

A. An attendant in a mental hospital
B. A person who worries about what others think of him
C. A person who enjoys being different from the majority
D. A person who enjoys playing jokes on others
E. An insane person

The answer had been selected so that one could clearly be best, while two would be irrelevant or contradictory to the sense of the quotation. Several sets of possible answers of this type were tried out on local teachers before the ones used on the test were chosen. Either of the worst answers was worth one point, a best answer was worth five points, and either of the other answers was worth three points. The range of possible scores was from 10 through 50.

The judgment-of-quality test utilized a technique that dates back at least to the 1920s. Ten brief passages from the works of noted writers were chosen. Each of them was rewritten in such a way as to introduce illogicalities and infelicities and then rewritten again to introduce even more inelegant touches, so that the third version was in effect a parody of the original. Among adult readers of these items there was 100 percent agreement as to which was the best and worst version. The following three versions of a stanza from a Longfellow poem were on one form of this test.

A.

Were all the guns, that fill the world with terror,
Were all the wealth, bestowed on politicians,
Given to cure the human mind of error,
There were not need of buying ammunitions.

B.

Were half the power, that fills the world with terror,
Were half the wealth, bestowed on camps and courts,
Given to redeem the human mind from error;
There were no need of arsenals and forts.

C.

Were half the power that fills the world with terror,
Were all the wealth that's stolen by politicians,
Used to free men from the burdens that they bear,
And to train scientists and technicians.

Students were asked to select both the best and the worst versions. The proper choice was worth two points, a second best choice worth one point. Scores could range between 0 and 40.

The Philosophical-Insights Category

Constructing an objective test that would measure changes in this area—one of great concern to English teachers, according to our earlier study—proved extremely difficult. The forming of judgments about student progress in such an area is a matter of observing the patterns of a student's utterances and behaviors over a considerable period of time. We settled for a test which attempts to get at the student's perceptions of the philosophical or ethical orientation of the author of the play as expressed in the particular work. Even at this, the questions we could devise were so complex that few of them could be used. There were ten items in the thematic-understanding test, each having the following form:

Consider everything that happens to Macbeth in the play—what he does, what he experiences, and what he may have learned from all of it. Then, imagine you are able to ask one question to Macbeth's ghost. Which of the three suggested answers do you think would come closest to the one Macbeth's ghost would give?

THE QUESTION: "Some people say that man's fate is determined by powers beyond his control, and other people say that everyone has control over his own fate and is responsible for what happens to him. What do you think?"

THE ANSWERS:

A. "I think that everything is predetermined and that no one has any control over what happens to him."
B. "A man is master of his own fate, and he must take the responsibility for what he does."
C. "I don't know. It's confusing. You'll have to find the answer for yourself."

We somewhat arbitrarily classified the answers to each question as most acceptable, possibly acceptable, and unacceptable. A most acceptable answer was worth two points and a possibly acceptable answer worth one point, so the scores could range from 0 to 20.

The Appreciation Category

Although almost everyone values appreciation as an outcome of experiences with the arts, there was no factor that emerged from our analyses of the data from the drama objectives study that could be associated with appreciation. The case seemed to be that appreciation was thought of either in connection with a specific art form, e.g., appreciation of literature, or that appreciation items were grouped with other objectives according to some set of not quite definable criteria. Examination of the items that had been assigned to the appreciation pool suggested that they might profitably be classified according to the mental operations involved. After a number of preliminary attempts at subclassification, we finally decided on three subcategories that distinguished (1) attitudes toward theatre, literature, and the arts; (2) cognitions about the nature, function, or power of the theatre, literature, and the arts; and (3) discriminative behaviors indicative of the internalization of the foregoing attitudes and cognitions.

The attitudes test consisted of thirty statements of attitudes toward the theatre or one of the arts. Twenty of the statements were phrased positively and ten negatively. The student was asked to express his agreement or disagreement with each statement. One of the statements on this test read:

It would be very exciting and stimulating to work in the theatre.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

Scoring was on the basis of four points for the most favorable answer through one point for the least favorable answer, giving a range of possible scores from 30 to 120.

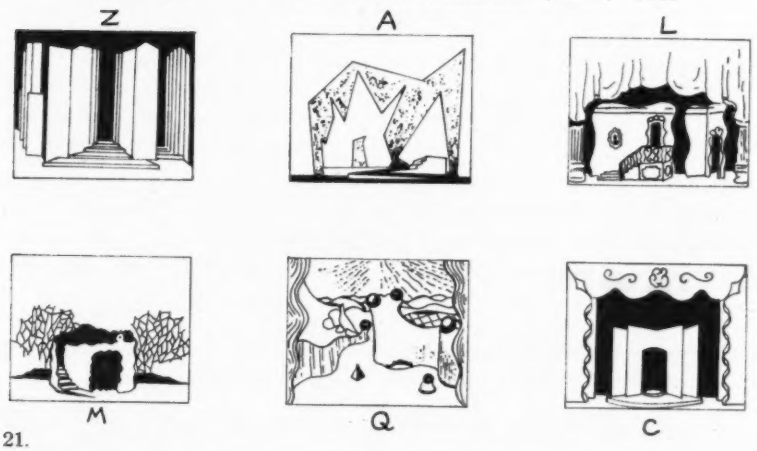
The cognitions test was constructed and scored in the same way as the attitudes test. A sample item read as follows:

Plays can make you care about things that never made any difference to you before.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

The discrimination test was frankly experimental. It consisted of six deliberately-rough drawings of a set on a proscenium stage (Figure 1). Ten simple plot outlines were written to describe various types of plays: farce, fantasy, realistic drama, tragedy, and so on. Some of the plots were adapted from classic

Figure 1
A Sample Item from the Discrimination (ADP) Test



21.

DIRECTIONS. The six sketches above represent stage settings for plays. Below is a plot outline of a play. Read the plot outline and decide which of the six settings would be most appropriate for the plot. On the answer sheet, find the letter that identifies the setting you have chosen and circle it. The letters on the answer sheet are not in the same order as the pictures in most cases. Please make sure you circle the letter that you intend to circle.

THE PLOT

The main characters in this play are two lonely and embittered old men, isolated from life and the world. They talk to one another and to characters who pass through about the emptiness of existence, about leaving the place where they are, and about doing something important. But at the end of the play they are still standing just where they were when the play opened, still lonely and still isolated.

plays; others were invented to be appropriate to one of the sets. The students' task, as explained in the directions in Figure 1, was to choose the setting most appropriate for a performance of the play described in the plot outline. Trying to take account of the difficulties of evaluating responses to a question such as this—e.g., a creative student might consciously choose the least appropriate set for its ironic effect—we classified the six sets in relation to each plot outline as most appropriate, possibly appropriate, and inappropriate. A most appropriate choice was worth two points and a possibly appropriate choice one point, so the range of possible scores was from 0 to 20.

The Desirable-Attitudes-and-Behaviors Category

The items assigned to this category involved social learnings from the theatre and their transfer to other situations, including the classroom. Because it was of special interest within the Project, a separate theatre-etiquette test was constructed. The desirable-attitudes test consisted of statements of changes in attitudes which had come about as a result of the experience of attending theatre. About half the items were phrased in the first person and half phrased as descriptions of what had happened to other students. The respondent was to express his agreement or disagreement with each statement, as in the following example.

> Being part of the audience at a live play has made me more aware of how important it is to listen carefully.
>
> A. Strongly agree
> B. Agree
> C. Disagree
> D. Strongly disagree

The twenty-item behaviors test was similarly constructed, but the statements had to do with changes in actual behavior as a result of experiences in the theatre.

> My class seems to listen better and to be more attentive after their theatre experiences.
>
> A. Strongly agree
> B. Agree
> C. Disagree
> D. Strongly disagree

The range of possible scores on the attitudes test was from 30 to 120; on the behaviors test it was from 20 to 80.

The theatre-etiquette test consisted of thirty statements, some phrased as reports of the respondent's in-theatre behavior and some phrased as reports of the behavior of other students. Again the respondent was to express agreement or disagreement with each item. The range of possible scores on this test was from 30 to 120.

Fewer students were impolite or inattentive at the play than in school.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

The Covariates Category

It seemed reasonable to believe that a student's intelligence would affect his performance and scores on such tests as those of interpretation and knowledge, and that the extent of his prior experience with the theatre and with dramatic activities would affect his responses on such tests as those in the desirable-attitudes-and-behaviors category. So it was necessary to take some account of these variables. As discussed on page 19, we could have entered these variables into the experimental design as treatment variables. For practical reasons, however, we used the verbal-intelligence and prior-theatre-experience scores as covariates, which is to say that before any other analyses were performed, calculations were made of the amount of variation in each dependent variable score that was attributable to verbal-intelligence and prior-theatre-experience scores. Then the mean scores on all the dependent variables were adjusted by that amount. So all of the mean scores reported hereafter are adjusted means which no longer reflect the influence of the verbal-intelligence and prior-theatre-experience measures.

The thirty-item verbal-intelligence test was constructed by sampling thirty items at random from a longer standardized test of verbal intelligence. The items were all of the analogies type, e.g., _____ is to man as fur is to _____, with the respondents being required to choose the pair of words from an accompanying set which best completed the analogy. The range of scores on this test was from 0 to 30.

The prior-theatre-experience test consisted simply of questionnaire-type items. The value of each response is noted in parentheses following the response. A respondent's prior-theatre-experience score was the sum of the values of the responses he chose. Scores on the verbal-intelligence and prior-theatre-experience variables were obtained before the start of the experiment.

Have you ever participated in putting on a play for an audience?

A. I have acted a major part (1)
B. I have acted a minor part (1)
C. I have been in a singing or dancing chorus (1)
D. I have worked on scenery, make-up or other backstage jobs (1)
E. I have worked as a ticket taker or usher at a play (1)
F. I have never done any work on a play (0)

Have you ever seen a live play in a theatre?

A. Yes, I have seen many plays (2)
B. Yes, I have seen one or two plays (1)
C. No, I have never seen a live play (0)

How many plays have you read or studied in your English classes?

A. Three or more (2)
B. One or two (1)
C. None (0)

The Validity of the Instruments: Some Comments

The power of any experimental design is ultimately a function of the quality of the dependent measures. If the instruments used to quantify the dependent variables are invalid, then the study will be of little value. In the areas of response to theatre and response to literature there has been very little previous work that is of high enough quality to be useful to a researcher. Therefore, one of our central concerns throughout the three years in which we were assessing the Educational Laboratory Theatre Project had to be the development of measuring instruments and techniques.

We have availed ourselves of the established techniques for measuring knowledge and attitudes, and we have used such instruments as the semantic differential. We have tried to get at such variables as student response to a theatrical production by a variety of methods: ratings by the actors, in-depth inter-

views with students, systematic and informal observations in the theatre, and the electronic recording of the volume of student responses at crucial points in a play. Some of the measuring techniques we have developed seem to hold promise, and they have been or will be reported on elsewhere.

But in general what we have found is that the techniques which seem best able to get at the internal responses of students are those which are clinical, rather than objective, and which, by their nature, are extremely time-consuming both to administer and to analyze. A projective test, for example, yields data which must be coded or content analyzed by a number of judges; and the development of a set of scoring protocols which will ensure acceptable interjudge reliability is a long and intricate task. Constraints are set upon the number of subjects and the number of variables that can be so examined by the time, money, and trained manpower that are available.

It is difficult to generalize convincingly from the clinical study of a small number of subjects. In addition, the number of independent variables which may be manipulated is restricted when the number of subjects is small, and the number of dependent variables which may be measured is reduced when the scoring procedures are time-consuming and expensive. So the time comes in the planning of an experiment when the researcher must decide whether it is more appropriate in a particular case to study a few subjects and a few variables intensively, using qualitative techniques, or to study a large number of subjects and a large number of variables using objective tests.

In other studies we had done we had chosen to use qualitative techniques; but in the present case, because the hypotheses at question were general statements of pedagogical theory which purportedly involved powerful factors and applied to students in the mass, it seemed appropriate to sacrifice depth for the sake of extensiveness.

As has been noted at length, the design of the present study controls for most of those extraneous factors which could influence scores on the dependent measures. But even in the present case the questions must be asked, when there are negative findings, whether the dependent measures were adequate. Did they really measure what they purported to measure? Were they sensitive enough to register differences which existed? So

there must be some discussion of the validity of the instruments used to define the dependent variables. But before proceeding, let us note that, regardless of how much some of the instruments used in this study might fall short of the ideal, all of them were much more carefully constructed than the tests which are used in the schools from day to day as the basis for decisions which will affect the lives of students and the fates of programs. That is to say, it would do no harm to consider the tests used in this study as superior versions of conventional teacher-made tests or as draft versions of standardized tests of the future.

Any researcher in an area such as that involved in the present study has little choice but to construct his own instruments as well as he is able. We would argue that each of the instruments used in this study does measure that property designated in its title, and we would admit that some of the dependent measures are probably more valid indicators than others of the sorts of behaviors that were referred to in speculations about the effects of different methods of preparing students for the theatre.

In particular, the involvement test obviously gets at only a tiny part of the complex of behaviors to which a theatre person is referring when he talks about students having an intense experience or being a good audience. Similarly, the thematic-understanding test certainly does not sample everything that English teachers are referring to when they speak of literary studies giving students ethical and philosophical insights. And the discrimination test is more or less an unknown quantity, an attempt to quantify an aesthetic judgment.

But a good argument can be made for the content validity of all of the other tests. (In the absence of both an adequate theory of literary response and any significant amount of empirical work, it is not worthwhile discussing the other sorts of validity.) The pools of items from which the test items were selected were very large; after the elimination of redundancies and mere verbal variations on the same item, each pool represented something as near to a population of possible instances of each property as we could contrive. Five qualified judges had agreed that the items in each pool were specifically representative of the property to be measured. And the items making up each test were randomly sampled from the larger pools of items.

The discriminating power of the tests cannot be demonstrated

except in those cases where statistically significant effects were found; but each of the tests yielded a wide range of class mean scores. And finally, although conventional measures of reliability cannot be computed because the tests are item-sampled, the means scores and ranges of scores between the two replications were quite comparable.

Other Measures

In addition to the tests described above there were a number of other pieces of information gathered that were external to the experimental design itself. A pretest instrument which was used to get verbal-intelligence and prior-theatre-experience scores also contained the following six statements:

I watch TV much less than I did six months ago.

Literature is the most important part of English.

There is no reason to discuss and analyze literature; we should just read it and enjoy it.

The most important thing about literature is that it tells us how to behave morally.

I can understand literature better if I read it aloud and act it out.

I read much more than I did six months ago.

The student was to express the degree of his agreement or disagreement with each statement: strongly agree, agree, don't know, disagree, strongly disagree.

These six statments were repeated on a questionnaire which was circulated to a sample of approximately 25 percent of the classes which had taken part in the experiment about a month after the completion of the last phase of the experiment. Our intention was to see what changes, if any, might have taken place in the areas touched on by these items during the entire course of the experiment. The results of these comparisons, not being directly relevant to this study, will not be reported here.

Other items included on the test instruments were intended to provide a check on the teacher's behavior, so that we might take account of any gross departures from a prescribed treat-

ment. The alternatives differed slightly for the two plays involved in the experiment. The items were these:

Have you seen the production of_____?

A. Yes
B. No

Have you read all or part of_____?

About how much time did your class spend in studying or discussing the Project Discovery production of_____or matters related to it? (Include in your estimate time spent studying other plays by_____, background materials, and drama in general; also include time spent out of class doing library research assignments. Do *not* include the time spent reading the play at home.)

A. Two hours or less
B. Between two and four hours
C. Between four and six hours
D. Between six and eight hours
E. More than eight hours

Of all the time spent in your English class studying matters related to the Project Discovery production of_____, approximately what fraction of time was devoted to having students read aloud from the play or act out scenes from it?

A. No time
B. One-fourth of the time
C. One-half of the time
D. Three-quarters of the time
E. Almost all of the time

As already noted the students' reports of teacher behavior merely served to confirm that the teachers were indeed doing what they had agreed to do, and no further use was made of the information produced by these items.

One final type of data not previously mentioned was also gathered. Thinking it was possible that effects of the different treatments might be manifested during the spontaneous discussions in the classroom immediately following the play, we decided to observe a number of classrooms in different treatment conditions. Approximately twenty classes were visited on the day after the students had seen the first play of the season. The observer, Phyllis Hubbell of the CEMREL staff, during each period

made three types of observations in successive five-minute blocks, so that in each class period two or three five-minute records of each type were obtained. In one five-minute block observations of teacher and student verbal behavior were made on a systematic observational schedule we had adapted from schedules developed by other researchers. In another five-minute block field notes were taken. And in the third five-minute block the content of the ongoing discussion was classified every thirty seconds to record whether it related to the performance, the text of the play, personal reactions, irrelevant matters, etc.

Analyses of the data showed differences between classes within a rather narrow range, but the differences had no systematic relationship to the experimental treatments given the various classes. Without promise of results, this type of data was too expensive to be gathered, and therefore the observations were not repeated in connection with the second play.

CHAPTER FIVE

THE PLAN FOR ANALYZING THE DATA

The experiment was designed so that a multivariate analysis of variance (MANOVA) could be used. Multivariate analysis of variance is a procedure by means of which two or more independent and dependent variables can be evaluated simultaneously. It is a method which has become practical to use only since computers have become readily available. However, now that MANOVA programs which will handle complex designs are in computer center libraries, the technique is available even to researchers who do not fully understand the mathematics of it.[1]

All it seems necessary to do here is to lay out the contrasts we examined and to comment on the peculiarities of the fractional factorial design which place restrictions on our interpretations of the contrasts.

The information in this chapter is not essential to an understanding of the results reported later. The chapter is intended primarily to acquaint the aspiring researcher with some of the ways in which one can handle data from an experiment such as this. Though it is too simplified to satisfy the methodologically sophisticated reader, it is probably too technical for the general reader to follow easily. Therefore it is suggested that the reader without a special interest in this part of the experimental design should turn ahead to Chapter Six whenever he finds himself becoming too confused to continue.

The Contrasts

Tables 9 through 15 present the scheme followed in the analyses of the data from this experiment. The whole series of analyses outlined in the tables was carried out for each hy-

[1] The MANOVA program we used was NYBMUL, written by Jeremy Finn, Department of Educational Psychology, State University of New York at Buffalo. We used the revision of the NYBMUL program dated June 19, 1969, and published by the Computing Center, State University of New York at Buffalo.

pothesis, i.e., for each independent variable and each combination of independent variables. In this and the following chapters the term hypothesis should be understood to refer to the question of whether a particular independent variable or combination of independent variables had significant effects. For example, the first hypothesis to be dealt with below is that the intensity of the study of background has an effect upon test scores.

The notational system used in the tables of contrasts is extremely efficient but it requires some explanation. The explanation will be easier to follow if it is given in terms of a set of data, and such a set of fictitious data is given in Table 8. The

Table 8

Mean Total Scores on All Dependent Measures at Two Levels of an Independent Variable (Fictitious Data for Illustrative Purposes)

Code Name of Dependent Variable	*Level of the Independent Variable* +	−
LIK	4	6
INV	4	4
NOQ	5	6
NOT	3	5
PHI	6	5
APA	4	4
APC	3	3
ADP	5	4
DAT	4	5
BEH	6	5
ETQ	5	4

scores entered in the columns of Table 8 represent mean total scores, which is to say that the LIK mean at the + level in Table 8 is the sum of the mean XLIK score plus the mean YLIK score for all classes at the + level of the independent variable in question. Since any class of students at the + level in the first block of the experiment would always be at the − level in the second block, the mean scores at both the + and − levels of the variable have been contributed by the same subjects.

Table 9

Structure of Contrasts for the MANOVA with the Eleven Student Performance Measures (Total Scores)

Dependent Variables	*Contrasts*										
	LIK	*INV*	*NOQ*	*NOT*	*PHI*	*APA*	*APC*	*ADP*	*DAT*	*BEH*	*ETQ*
LIK	1	0	0	0	0	0	0	0	0	0	0
INV	0	1	0	0	0	0	0	0	0	0	0
NOQ	0	0	1	0	0	0	0	0	0	0	0
NOT	0	0	0	1	0	0	0	0	0	0	0
PHI	0	0	0	0	1	0	0	0	0	0	0
APA	0	0	0	0	0	1	0	0	0	0	0
APC	0	0	0	0	0	0	1	0	0	0	0
ADP	0	0	0	0	0	0	0	1	0	0	0
DAT	0	0	0	0	0	0	0	0	1	0	0
BEH	0	0	0	0	0	0	0	0	0	1	0
ETQ	0	0	0	0	0	0	0	0	0	0	1

See page 58 for explanation of Tables 9–15.

Table 10

Structure of Contrasts for the MANOVA with the Four Tests in the Affective-Response Category

	XLIK	*YLIK*	*XINV*	*YINV*	*Means*	*X – Y*	*LIK – INV*	*LIKINVXY*
XLIK	1	0	0	0	1	1	1	1
YLIK	0	1	0	0	1	−1	1	−1
XINV	0	0	1	0	1	1	−1	−1
YINV	0	0	0	1	1	−1	−1	1
		Set 1					Set 2	

Table 11

Structure of Contrasts for the MANOVA with the Four Tests in the Knowledge Category

	XNOQ	*YNOQ*	*XNOT*	*YNOT*	*Means*	*X – Y*	*NOQ – NOT*	*NOQNOTXY*
XNOQ	1	0	0	0	1	1	1	1
YNOQ	0	1	0	0	1	−1	1	−1
XNOT	0	0	1	0	1	1	−1	−1
YNOT	0	0	0	1	1	−1	−1	1
		Set 1					Set 2	

Table 12

Structure of Contrasts for the MANOVA
with the Two Tests in the Philosophical-Understandings Category

	XPHI	*YPHI*	*Means*	*X − Y*
XPHI	1	0	1	1
YPHI	0	1	1	−1
	Set 1		Set 2	

Now refer to Table 9, which summarizes the contrasts between total test scores that were actually examined under each hypothesis. Each row in the matrix designates a dependent variable or test according to the labels at the left. Each column describes a contrast. The first column in Table 9 is headed LIK, and the column consists of a 1 in the LIK row and zeroes in all other rows. The 1s and 0s are weights, and the column indicates that in computing the LIK contrast the observed mean scores at each level of the independent variable are to be multiplied by the designated weights. It had already been noted that the + and − symbols used to designate levels of the independent variables are also in fact weights, namely, +1 and −1.

What the first column in Table 9 designates then is a series of operations to be followed in order to obtain the difference score which is to be tested for significance. Referring to the fictitious data in Table 8 we find that the mean total score on the LIK test is 4 at the + level and 6 at the − level. Multiplying these by the weights and summing gives us $+1(4) - 1(6) = -2$. The mean total scores on each of the other tests are treated in the same way, and then the sum of each of these pairs of scores is multiplied by the weight designated in the LIK column of Table 9 and all of the scores are summed. The column sum is the score to be tested for significance. Using the scores in Table 8 these operations would yield:

LIK	*INV*	*NOQ*	*NOT*		*ETQ*	
Scores	*Scores*	*Scores*	*Scores*	. . .	*Scores*	
$1(4-6)$	$+0(4-4)$	$+0(5-6)$	$+0(3-5)$	$+ \ldots +$	$0(5-4)$	$= -2$

Since scores that are multiplied by the weight zero are in effect eliminated, the LIK column is simply a way of asking

Table 13

Structure of Contrasts for the MANOVA with the Six Tests in the Appreciation Category

	XAPA	*XAPC*	*XADP*	*YAPA*	*YAPC*	*YADP*	*Means*	*X − Y*	*APA − ADP*	*APAADPXY*	*APA − APC*	*APAAPCXY*
XAPA	1	0	0	0	0	0	1	1	1	1	1	1
XAPC	0	1	0	0	0	0	1	0	0	0	−1	−1
XADP	0	0	1	0	0	0	1	−1	−1	−1	0	0
YAPA	0	0	0	1	0	0	1	1	1	−1	1	−1
YAPC	0	0	0	0	1	0	1	0	0	0	−1	1
YADP	0	0	0	0	0	1	1	−1	−1	1	0	0
			Set 1						Set 2			

	Means	*X − Y*	*APA*	*APC*	*ADP*	*APCADPXY*
XAPA	1	1	1	0	0	1
XAPC	1	1	0	1	0	0
XADP	1	1	0	0	1	−1
YAPA	1	−1	1	0	0	−1
YAPC	1	−1	0	1	0	0
YADP	1	−1	0	0	1	1
				Set 3		

Table 14

Structure of Contrasts for the MANOVA with the Six Tests in the Desirable-Attitudes-and-Behaviors Category

	XDAT	*XBEH*	*XETQ*	*YDAT*	*YBEH*	*YETQ*	*Means*	*X – Y*	*DAT – ETQ*	*DATETQXY*	*DAT – BEH*	*DATBEHXY*
XDAT	1	0	0	0	0	0	1	1	1	1	1	1
XBEH	0	1	0	0	0	0	1	1	0	0	−1	−1
XETQ	0	0	1	0	0	0	1	1	−1	−1	0	0
YDAT	0	0	0	1	0	0	1	−1	1	−1	1	−1
YBEH	0	0	0	0	1	0	1	−1	0	0	−1	1
YETQ	0	0	0	0	0	1	1	−1	−1	1	0	0

Set 1 Set 2

	Means	*X – Y*	*DAT*	*BEH*	*ETQ*	*BEHETQXY*
XDAT	1	1	1	0	0	0
XBEH	1	1	0	1	0	1
XETQ	1	1	0	0	1	−1
YDAT	1	−1	1	0	0	0
YBEH	1	−1	0	1	0	−1
YETQ	1	−1	0	0	1	1

Set 3

Table 15

Structure of Contrasts for the MANOVA with the Two Tests in the Interpretive-Skills Category

	ADW	*INT*
YJUD	1	0
YINT	0	1

whether, under the particular hypothesis, LIK scores differ significantly between the two levels of the independent variable.

The second row is headed INV and consists of zeroes except for a 1 in the INV row. The type of analysis we used is called a step-down analysis, which means that as each analysis in a series is performed, the portion of the total variance attributable to the variable being evaluated is taken out. So the second column is a way of asking whether, under the particular hypothesis, there are differences in INV scores after variance due to LIK scores is removed. The third column asks whether there are differences between NOQ scores after variance due to both LIK and INV is removed.

Tables 10 through 15 summarize the analyses of the tests within the categories of dependent variables discussed in Chapter Four. A consideration of one of these sets of analyses should make clearer the principles on which our treatment of the data were based. Table 10 is devoted to the affective-response category, a category made up of four tests—the liking tests for the first and second replications of the experiment (XLIK and YLIK) and the involvement tests for the first and second replications (XINV and YINV). Two sets of contrasts are summarized in the table. Each of the sets represents a different way of partitioning the total variance. The four contrasts in Set 1 in Table 10 partition the variance by forms of the tests. In Set 2 the variance is differently partitioned; in effect, Set 2 represents a reconceptualization of the variables making up the category, or the creation of a new set of dependent variables. The reason for the creation of new scales is to seek the best, i.e., the most parsimonious, explanation of what significant effects may be found.

The first column in the second set is headed means. It is conceivable that an effect of an independent variable might be to inflate the general level of mean scores at one level on all

tests. Assume that the total LIK and INV scores in Table 8 were the sums of the following mean scores on the individual tests:

	+	−
XLIK	2	4
YLIK	2	2
XINV	3	2
YINV	1	2

The 1s in each row of the means column in Table 10 would call for the following operations:

$$1(2-4)+1(2-2)+1(3-2)+1(1-2)=-2$$

Such a result would indicate that at one level of the independent variable in question the effect was to inflate the general level of mean scores. This difference would be tested for significance, and the portion of the total variance due to differences between means would then be taken out before the next contrast was evaluated.

For the sake of simplicity, the step-down feature of the analysis will be ignored for the moment and the other contrasts in the set will be gone through using the data from Table 8, so that the notational scheme may be thoroughly clarified. The second column of Set 2, Table 10 is headed X − Y. The operations prescribed in the column evaluate the differences between the summed scores on the two tests for the first play and the summed scores on the two tests for the second play. Multiplying the differences between mean scores by the designated weights and summing down the column would give us:

$$1(2-4)-1(2-2)+1(3-2)-1(1-2)=0$$

This result would indicate that there were no differences between blocks in scores on tests in the affective-response category.

The third column is headed LIK − INV. It evaluates the difference between the summed LIK scores and the summed INV scores. The operations called for in the column would give us:

$$1(2-4)+1(2-2)-1(3-2)-1(1-2)=-2$$

For this data the result would indicate that there were differences in the way that the independent variable affected total scores on the two tests. The final column headed LIKINVXY evaluates the interaction between tests and occasions and calls for the following operations:

$$1(2-4) - 1(2-2) - 1(3-2) + 1(1-2) = -4$$

This figure would estimate the portion of the total variance that might be explained in terms of the relationships between the tests defining the category and their interactions with the plays, performances, and so on, which differentiate one block of the experiment from the other.

To summarize, the matrices in Tables 9 through 15 lay out the analyses to which the data were subjected. The whole series of analyses was conducted for each of the hypotheses. Each matrix represents a way of partitioning the total variance in the test scores in question. Each column in a matrix represents a particular question asked of the data; the figures in each column are weights to be applied to the mean scores associated with the variables named in each row of the matrix. So each column may be taken as a description of the operations that are to be carried out in order to answer a particular question.

To go a step further, each of the matrices describes analyses to be made on the set of scores on the tests which identify the rows of the matrix. There is a certain amount of variance associated with each set of scores, and this amount may or may not be significantly different from zero. An F-ratio test of equality of mean vectors was used to establish whether the variance within each set of scores was significant.

Normally there is no point in further examining differences within a set of scores when the total variance associated with the scores is nonsignificant. However, in respect to the analyses of total scores on all eleven dependent measures (Table 9) there are two reasons why this criterion does not apply. First, when a step-down analysis is being used, the ordering of the variables is of crucial importance, since that portion of the variance which is not attributable to the independent variable becomes a proportionately larger part of the remaining variance with each successive analysis. In the cases of the tests grouped within categories we had fairly good reasons for arranging the tests in

particular orders. But in the case of the whole set of eleven total scores we had no such grounds for putting the tests in a certain order. Second, a number of the tests, especially those concerned with the transfer of learning, seldom or never discriminated between treatment conditions, probably because the behaviors in question are changed over a longer period of time than that covered by this study. At any rate, the inclusion of a number of such tests would reduce the total variance associated with the whole set of tests. Therefore, in regard to the tests of differences between total mean scores we were guided in our reporting not only by the obtained step-down F-ratios, but also by the univariate F-ratios, i.e., those computed independently of all other scores.

Analyses of Effects and Interactions

It was noted earlier that each column in a matrix was a way of asking the question whether, under a particular hypothesis, there were differences between the scores on a test at different levels of the independent variable in question. Fifteen hypotheses about each test or category of tests were evaluated, although only ten of these are strictly interpretable. Four of these hypotheses involved the effects of a single independent variable, and in such cases one speaks of evaluating the main effects of the variable. The other hypotheses involved two or more independent variables, and in these cases one speaks of evaluating interactions.

The available hypotheses involve main effects, two-factor interactions, three-factor interactions, and so on. But as noted earlier, when a fractional replication of a factorial design is used so that the number of runs will be reduced, one of the consequences is that certain effects are confounded with others. Two effects are confounded when a single set of computations is used to estimate an effect which may be interpreted as due to any one of two or more factors. In this design main effects are confounded with four-factor interactions, e.g., A with BCDE, and two-factor interactions are confounded with three-factor interactions, e.g., AB with CDE, according to the pattern shown in Table 16. The effects confounded with the effects in which we are interested are technically referred to as aliases. Each effect is ascribed to the factor or interaction in the hypothesis and to

Table 16

Summary of the Hypotheses Evaluated, Plus Other Possible Contrasts and the Alias Structure *

Hypothesis (Source)	*Alias*
1. A (background)	BCDE
2. B (text)	ACDE
3. C (timing)	ABDE
4. D (materials)	ABCE
5. AB (background × text)	CDE
6. AC (background × timing)	BDE
7. AD (background × materials)	BCE
8. BC (text × timing)	ADE
9. BD (text × materials)	ACE
10. CD (timing × materials)	ABE
ABC (background × text × timing)	DE
ABD (background × text × materials)	CE
ACD (background × timing × materials)	BE
11. BCD (text × timing × materials)	AE
ABCD (background × text × timing × materials)	E

* Only the numbered hypotheses are discussed in this report.

its alias. A good rule to follow in working with this sort of analysis is always to prefer the simpler explanation of a significant result. This means that if the AB effect is significant and the AB is confounded with CDE, we would ascribe the effect to the two-factor rather than the three-factor interaction.[2] The three-factor interactions in the first column of Table 9 have two-factor aliases. But one of the factors in each of the two-factor aliases is variable E, play performance, a single level of which

[2] Three-factor interactions have rarely been found to be significant; usually they make less conceptual sense than main effects or two-factor interactions. Edwards, in the following passage, speaks of the assumption that higher order interactions are "negligible": "If we use a ½ fractional replication of a 2^5 design, then each main effect will be confounded with a four-factor interaction. For example, the main effect of A will be confounded with B × C × D × E. Each two-factor interaction will be confounded with a three-factor interaction. For example, A × B will be confounded with C × D × E. If we can assume that all four- and three-factor interactions are negligible, then a ½ fractional replication of the 2^5 factorial experiment will provide information about all of the main effects and also about the two-factor interactions." (Reprinted with permission from *Experimental Design in Psychological Research* by Allen L. Edwards, published by Holt, Rinehart and Winston, Inc. Copyright © 1968. [pp. 256–257])

is common to all treatments. The interactions involving variable E do not therefore make good conceptual sense.

The design is not a satisfactory one for evaluating three-factor interactions, and we may therefore attend only to the four main effects and six two-factor interactions in the first column of Table 9. We will make one exception to this, however, in the case of the BCD interaction, because one of the hypothses ascribed to English teachers was that intensive study (B) of the play (D) should take place before the performance (C).

CHAPTER SIX

OTHER FEATURES OF THE STUDY

Item Sampling

In reading the section on the tests that were used as dependent measures in this study it must have occurred to the reader that the administration of all those tests would be so time-consuming as to interfere not only with the orderly conduct of the experimental classes but with the experiment itself. Actually, the total amount of each student's time that was devoted to test-taking amounted to perhaps an hour and a half spread over five testing periods.

We used what are known as item-sampling procedures to construct our data-gathering instruments. Item-sampling is a technique in which all the items on a test are randomly divided into a number of nonoverlapping samples. Each student in a class will answer only the fraction of the test items in one particular sample. In the present case, all of the tests that had ten or more items were item-sampled.

With a thirty-item test, three items were assigned to each of ten forms of the test. Within each experimental class the forms were randomly distributed. In a class of thirty students three students would take each form of the test. The mean scores of each set of three students responding to the same set of items would be computed, and the sum of the ten sets of mean scores would represent the mean score for the class on the test.

A form of item-sampling is being used in the National Assessment study, and the technique has the obvious advantage of allowing the researcher to get a great deal of information in a very short time. The technique is also very economical from the point of view of the time and money it takes to score the tests.

With a test made up of binary items, e.g., a true-false test, it is well established mathematically that item-sampling gives a better estimate of the true mean score—the one that would

be obtained if every student took the entire test—than any other method of sampling.[1]

Most of the tests we used, however, were not made up of binary items, and there is no explicit theoretical rationale for item-sampling from such tests. We resorted therefore to two sorts of empirical checks upon our procedures. First, we administered all the items in two of the tests to all students in the experimental classes in one school. The class means obtained in this way were compared with the means obtained earlier using item-sampling procedures, and the difference between the two sets of mean scores was smaller than one might have expected to find in a test-retest situation using a single method of administration. Second, we administered several entire tests to classes not involved in the study. Scoring only three designated items from each respondent's test created a simulation of the item-sampling situation. This procedure was repeated several times, using a series of different assignments of subjects to forms, and the series of class means obtained this way were compared with the actual class means. It will be sufficient to note here that the results of these empirical checks gave us confidence in the item-sampling procedures we were using.

It should perhaps be emphasized that the basic data in this experiment were class mean scores. One consequence of using the item-sampling technique as we used it is that nothing may be said about the scores of any individual student. That is to say, the subjects in the experiment were the fifty-two tenth-grade English classes, not the 1300 or so students in those classes. The mean of the mean scores of all the classes assigned to a particular level of an independent variable was the score that entered into the calculations to determine the significance of treatment effects.

Samples of the instruments created by use of the item-sampling procedure as well as a key explaining how items from the several tests were distributed on the instruments may be found in Appendix A of this paper.

---

[1] See the discussion of item sampling in Frederic M. Lord and Melvin R. Novick, *Statistical Theories of Mental Test Scores* (Reading, Mass.: Addison-Wesley Publishing Co., 1968), pp. 252–260. See also T. R. Husek and K. Sirontik, *Item Sampling in Educational Research* (Los Angeles: Center for the Study of Evaluation, 1967).

Assignment of Subjects to Treatments

We wanted to have at least two classes in each of the experimental conditions. It seemed wise to start out with a number of classes considerably larger than the desired minimum to give a margin for error and for attrition. Fifty-three teachers actually began the experiment, so that there were four randomly assigned classes in treatment conditions one through five and three classes in all other conditions. One of the teachers found it impossible to continue in the study and withdrew his class. Several others, because of schedule changes in the course of the first play, found that certain circumstances—e.g., too little time to complete an intensive treatment before the students attended the play—required that they be reassigned to another treatment condition.

For one reason or another we did not receive complete data from two of the classes. The design of the experiment, and the limitations of the computer program we were using, made it difficult to use anything less than a complete set of test scores. We decided it would be better to discard the data from these two classes than to estimate the missing scores. So the final number of teachers and classes contributing data to the study was fifty. After the necessary reassignments the fifty classes were distributed across experimental treatments as follows:

Run No.	*No. of Classes*	*Run No.*	*No. of Classes*
1	3	9	3
2	3	10	2
3	4	11	4
4	3	12	3
5	2	13	4
6	3	14	5
7	4	15	2
8	3	16	2

Three Uncontrolled Sources of Variation

Three extraneous factors were not taken account of in the design for this experiment, although there was reason to think that each of them, and the interactions between them, might possibly affect the scores on the various tests. The first, and probably least important, was the sequence of presentation of

the two plays and the two classroom treatments. *Red Roses for Me* was the first live stage play that most of the students in the experimental classes had ever seen. By the time these same students saw *Macbeth*, they may have been thinking, perceiving, and behaving in slightly different ways simply because they were now somewhat more sophisticated about theatre. So there may have been some sort of interaction between the experimental treatments and the sequence of presentation of the treatments. But there was no way in which we could have arranged to send students to see *Macbeth* first, so as to be able to estimate the sequence effects. In this case, circumstances made it impossible for the designer of the experiment to take into account a possibly noteworthy factor.

The other two uncontrolled sources of variation were the plays and the productions of the plays. The decision not to control for these factors was a deliberate one dictated not by circumstances but by the feeling that any available method of distinguishing levels of these variables would be so arbitrary as to be irresponsible and that the apparent advantages to be gained from typifying the plays would be spurious.

The design specialists whom we consulted were of the opinion that the design could be much neater if we could identify the two levels of the play variable as, for instance, tragedy and tragicomedy or Elizabethan tragedy and modern tragedy and the levels of the production variable as, for instance, conventional and unconventional, perhaps using the two factors as independent variables. Doing this would enable us to estimate play and performance effects. Or the two blocks might have been identified as conventional modern and unconventional Elizabethan. Then, if the effects under a particular hypothesis were significant for the X forms of certain tests but not for the Y forms, or vice versa, we might want to generalize from our findings to suggest that a factor had such-and-such effect in conjunction with a conventional production or a modern tragedy but another effect in conjunction with an experimental production of a Shakespearean tragedy.

We resisted this advice because it seemed to us that reifying such mere labels would tend to trivialize the whole study. To rephrase a familiar dictum in experimental terms, there are as many levels of the play factor as there are plays, and there are

as many levels of the performance factor as there are performances.

It seemed more responsible to us to consider each play and each performance as a unique event and to refrain from trying to generalize beyond the experimental situation itself in regard to the play and performance factors. Instead, we will discuss the important similarities and differences between the two plays and the two performances and leave it to the reader to generalize if he wishes. In any case, the sophisticated reader would reject an attempt to generalize from one production of *Macbeth* to Shakespearean plays in general or tragedies in general. And the less sophisticated reader would, unless specifically warned against it, tend to overgeneralize the results no matter how they were presented.

We hasten to add that it does not follow from the fact that each work of art is unique that scientific research in the arts is impossible. Let us imagine we could construct a huge matrix with all possible categories of plays arranged along one axis, and with all possible categories of performance styles arranged along the other. With 10,000 categories each of plays and performances, we would have a matrix of a billion cells, and the present experiment would enable us to say that certain things are true of the combinations of play and performance categories represented by two of these cells, and perhaps to guess that the same things were true of similar combinations.

We will in this case know more than we knew before we undertook the experiment. If enough additional repetitions of the experiment were to be conducted, at some point we should be able to hazard the theory that this or that is true of most possible combinations of plays and performances. This point would come the sooner, the better we were able to describe just how plays and performances may be classified and differentiated. But there is no easy way to shorten the process, and it would profit no one to pretend that presently available taxonomies of drama are scientifically adequate. Generalizations based on this assumption would give merely the illusion of yielding additional information.

We have gone into this matter at length because it seems to us that the problem of the uniqueness of art objects is one that faces anyone who wishes to do empirical research on teaching and learning in the arts, and we want to urge resistance to

the temptation to oversimplify the problem of dimensionalizing artistic stimuli, whether the simplification is undertaken for the sake of convenience or for the purpose of making one's work appear more important than it can possibly be.

To be sure one can get more attention by saying, "This is true about the teaching of drama," than he can by saying, "This is true of the teaching of these two plays in this time and place in conjunction with that particular pair of productions." But statements of the former sort are, today at least, scientifically irresponsible and certain to be discredited by later studies which will produce contradictory results. The generalizability of the results of an experiment such as the present one is not a statistical question, but an aesthetic and, in the broad sense, a strategic one. We are simply recommending modesty and the ability to be satisfied with small but certain gains in knowledge.

That position having been stated, let us examine some of the more important features of the two plays and the two productions involved in this study. Sean O'Casey's *Red Roses for Me* and Shakespeare's *Macbeth* have in common that they are generally considered too difficult for tenth graders. *Macbeth* is usually reserved for twelfth grade, and even the publisher of the paperback edition of *Red Roses for Me* advises English teachers that the play is suitable only as supplementary reading for gifted students. The experiences of the teachers in this experiment suggest that at least when live performances of the plays are available these estimates are far too pessimistic and that even below-average tenth graders can cope with either play.

The difficulties students have with Shakespeare's verse are legendary; but O'Casey makes demands upon his audience at least as great as those made by Shakespeare. O'Casey is the most lyrical of modern playwrights and the most nearly Elizabethan in the sweep and the extravagance of his language. Furthermore, both plays deal with issues and places unfamiliar to most students; if anything, the motivations of O'Casey's Dubliners are more obscure to Americans than those of Shakespeare's Scotsmen. Consider the following passages from *Red Roses for Me:*

> Ayamonn: Go an' lie down, lady; you're worn out. Time's a perjured jade, an' ever he moans a man must die. Who through every inch of life weaves a patthern of vigour an' elation can never

> taste death, but goes to sleep among the stars, his withered arms outstretched to greet th' echo of his own shout. It will be for them left behind to sigh for an hour, an' then to sing their own odd songs, an' do their own odd dances, to give a lonely God a little company, till they, too, pass by on their bare way out. When a true man dies, he is buried in th' birth of a thousand worlds.
>
> Finnoola: What would a girl, born in a wild Cork valley, among the mountains, brought up to sing the songs of her fathers, what would she choose but the patched coat, shaky shoes, an' hungry face of the Irish rebel? But their shabbiness was threaded with th' colours from the garment of Finn Mac Cool of the golden hair, Goll Mac Morna of th' big blows, Caolite of the flyin' feet, and Oscar of th' invincible spear.[1]

Thematically, both plays are concerned with civil conflict, fate, love, and ambition; and both end with the death in battle of the central character. But Macbeth's death restores the appointed order, while Ayamonn is a martyr in an unsuccessful demonstration against the oppressors of his people. Both plays are tragedies with touches of comedy, though there is certainly more of the latter in the O'Casey play. But the point is that this list of comparisons could be indefinitely extended without helping us to place the two plays in contrasting categories that have any real meaning.

This is even more true of the comparisons that can be made between the two productions. Both were done by the same artistic director and by the same repertory company. Both were polished professional productions in all respects. But *Red Roses for Me* was done on a proscenium stage with naturalistic settings and, except in the vision-of-Dublin interlude, naturalistic acting. On the other hand, *Macbeth* was played on an acting area that featured a board runway down the center of the audience and a multi-leveled scaffolding that surrounded the audience on three sides. The acting was stylized and the movement was fast-paced and elaborately choreographed. There were constant and ingenious uses of special effects of all kinds. Watching this *Macbeth*, which the critics variously termed "total theatre," "neo-Elizabethan," and "Macbeth in the Wild, Wild West," was a

[1] Reprinted with permission from *Three More Plays* by Sean O'Casey, published by St. Martin's Press, Inc., Macmillan & Co., Ltd. Copyright © 1965.

radically different experience from watching *Red Roses for Me*. But it was beyond our ingenuity to typify the differences in a way that would make meaningful generalization possible.

So the case is this. The design we utilized reduced the number of identified sources of uncontrolled variation to three, the first of which is probably insignificant. The two remaining potentially important sources of variation, the plays and the productions, are phenomena that are, in our present ignorance, simply too complex to be handled. These three factors contribute in some unknown way to the total variance, and the influence of any one of the factors must simply remain a subject for speculation. On the whole, however, there is little in the data to be reported later which suggests that the sequence, play, and production factors seriously affected the results.

CHAPTER SEVEN

PRESENTATION OF RESULTS

Summary of Significant Contrasts

In Table 17 the eleven tests administered during both replications of the experiment are listed in the first column. In the second column of the table are summarized the independent variables which had effects that reached the .05 level of significance. What is perhaps most notable about this summary is the relatively small number of significant contrasts. The experiment was carried out because experienced professionals in education and theatre were strongly of the opinion that student responses to the Theatre Project would be affected in important ways by variations in methods of treating the plays in the classroom.

But the timing of the classroom study, before or after the performance, had no significant effect on the scores on any of the tests. The content of the lessons, the performed play or a related one, significantly affected scores only on the knowledge and thematic-understanding tests. The intensity of the study of the text, brief or intense, significantly affected scores only on the appreciation: attitudes test; and, rather surprisingly, the background factor, brief or intense, figured in all of the significant interactions.

The third column in Table 17 summarizes the independent variables which had effects significant between the .05 and .10 levels. Except in a few cases these effects are not discussed, but the summary in the second column demonstrates that even if the criterion for significance were relaxed to .10 the pattern of the findings would not be drastically changed: the significant effects would still be relatively few; there would still be no significant main effects of timing; and the interactions between the factors would still be the most prominent source of significant effects.

Table 18 summarizes the independent variables which had significant or near-significant effects upon scores within the six categories into which the tests were grouped. The picture here

Table 17

Summary of Significant One- and Two-Factor Effects on Total Test Scores

Test Names and Code Designations	*Factors and Interactions Significant at .05*	*Significant between .05 and .10*
Liking (LIK)	None	Background × Timing
Involvement (INV)	None	Text
Knowledge: quotations (NOQ)	Content	Text × Content
Knowledge: true-false (NOT)	Background × Text	Background × Content
Appreciation: attitudes (APA)	Text; Background × Text	None
Appreciation: cognitions (APC)	Background × Timing	Background; Text; Content; Timing × Content
Appreciation: discrimination (ADP)	None	Timing × Content
Desirable attitudes (DAT)	None	Background; Text × Content; Timing × Content
Desirable behaviors (BEH)	None	Background × Text
Theatre etiquette (ETQ)	None	Text × Timing
Thematic understanding (PHI)	Content; Background × Text	None

Table 18

Summary of Significant One- and Two-Factor Effects upon Test Scores within Categories

Category Names	*Factors and Interactions Significant at .05*	*Significant between .05 and .10*
Affective response	Timing	Content; Background × Timing
Knowledge	Timing; Content; Background × Text	None
Interpretive skills	None	Timing; Text × Content
Philosophical insights	Content	Content
Appreciation	None	None
Desirable attitudes and behaviors	None	Timing; Content

differs from that given in Table 17 primarily in that (1) the significant effects are even fewer, but (2) they include significant main effects of the timing factor upon scores in the knowledge and affective-response categories.

Significant Findings Under Each Hypothesis

Only those effects which are significant beyond (or in some cases near) the .05 level are discussed in the sections below. For the reader interested in the detailed results of the analyses the tables in Appendix D summarize the F-ratios and significance levels for all total test scores under each hypothesis and for all within-category scores under each hypothesis, as in Tables 9 through 15 in Chapter Five.

In this part of the chapter, a section is devoted to each independent variable, i.e., to the four primary factors, the six two-factor interactions, and the BCD interaction. More properly, a section is devoted to each hypothesis that a particular independent variable had significant effects. Within each section attention is first paid to contrasts between total test scores, the analyses described in Table 9. F-ratios and mean scores are presented for significant effects, and the observed significant differences are discussed and interpreted.

Then, in each section attention is given to significant effects upon scores within categories. F-ratios and mean scores are given for these categories, and the results of analyses of the contrasts involving alternative conceptualizations of the dependent variables within the categories are presented when they help to explain the significant effects.[1]

Hypothesis 1: Intensity of the Study of Background

There were no significant main effects of the background factor, so that insofar as total scores on the tests are concerned the effects upon student performance of a brief study of the background of a play were indistinguishable from the effects of an intense study. In two cases background effects approached significance. On both the appreciation:cognitions test ($F_{1,32} =$

[1] Alternative conceptualizations refers to those contrasts in the second and third sets in Tables 10 through 15 which involve partitioning the total variance in other ways than by tests, e.g., between plays, between summed scores on the various tests within the category, and so on.

3.09; $P < .09$) and the desirable-attitudes test ($F_{1,32} = 3.62$; $P < .07$), it is interesting to note that the higher mean scores were associated with the brief study of the background.

Level of Intensity of Study of Background	*Mean Scores*	
	APC	*DAT*
Brief	189.5	175.8
Intense	186.8	170.8

This suggests that there is a point of diminishing returns when it comes to the intensity of study, and in the data to be presented below statistically significant evidence of this phenomenon will be presented. There were no significant or near-significant main effects of the background factor upon scores within any of the categories of tests.

Hypothesis 2: Intensity of the Study of the Text

The only significant main effect of the text factor was upon scores on the appreciation:attitudes test ($F_{1,32} = 5.77$; $P < .02$). The higher mean scores on this test are associated with the brief level of the factor.

Level of Intensity of Study of Text	*Mean Scores*
	APA
Brief	191.2
Intense	188.3

None of the effects of the text factor upon scores within the categories of tests approaches significance, so, except in the case of the appreciation:attitudes test, the effects of one or two periods of study are indistinguishable from the effects of from four to seven periods of study. This finding, which is confirmed several times in analyses reported later, suggests that when a performance is available, an adequately thorough study of a play need not consume so much time as to create problems for a teacher who feels pushed to cover the material in the curriculum.

Hypothesis 3: The Timing of the Classroom Treatment

None of the main effects of the timing factor upon total test scores approached significance. But when the categories of tests

were considered, there were two significant main effects of timing. Within the affective-response category ($F_{4,28} = 3.07$; $P < .03$), the timing of the lessons affected scores primarily on the two liking tests.

Test	$F_{1,28}$	*P*
XLIK	3.65	0.07
XINV	1.61	0.21
YLIK	6.05	0.02
YINV	0.32	0.58

But the differences in liking scores were in opposite directions for the two plays:

	Mean Scores	
Level of Timing	*XLIK*	*YLIK*
Before	4.17	4.23
After	4.02	4.46

The liking and involvement tests were administered immediately after each class had attended the performance, so the classes at the after level of the timing factor had had no classroom treatment at all before they judged the performance. In the case of the first play, *Red Roses for Me*, these after-students judged the performance less favorably than those who had received some preparation; but in the case of the second play, *Macbeth*, they judged the play significantly more favorably than students who had been prepared for it.

According to the data, the timing of the preparation affected the students' expressed liking for the play, but did not affect their reported involvement with it. The significant LIKINVXY interaction (XLIK − YLIK − XINV + YINV; $F_{1,32} = 7.31$; $P < .01$) may be taken as strengthening the interpretation that an interaction between the timing of the classroom preparation and the play and/or production of the play affected liking scores. The one highly significant difference between YLIK scores would support the actors' contention that students will enjoy plays more if they go to the theatre without preparation. The almost significant effect on XLIK scores supports the educators' contrary assertion. All of this suggests that it is unwise to state the question, How should students be prepared for plays?, in ab-

solute terms and that one must specify what sort of play and production should be prepared for or not prepared for.

As a start in this direction, a combination of data and external evidence gives grounds for suggesting tentatively that preparing students for a conventional production of a play may facilitate their enjoyment of it, while such preparation may inhibit student enjoyment of a total-theatre production of the play. Certainly it is not unreasonable to suggest that any sort of conventional classroom preparation might interfere with a student's response to the *Macbeth* which Adrian Hall mounted; it featured real cannons, a pansy witch, tympanies, apparitions descending from the rafters, very red blood everywhere, a belching porter, a light show, Macbeth swinging through the scaffolding to escape Macduff, and, to cap it, Macbeth's bleeding head on a pike paraded through the audience.

Within the knowledge category also there were significant timing effects ($F_{4,29} = 3.85$; $P < .01$). But by far the largest part of the variance was due to between-level differences on the first true-false test of knowledge (XNOT).

Test	$F_{1,a}$	*P*
XNOQ	1.04	0.32
XNOT	13.95	0.001
YNOQ	0.06	0.81
YNOT	0.59	0.44

The NOT tests, it will be remembered, consisted of forty play-specific true-false items dealing with facts about the plot and characters in each play. The common sense expectation would certainly be that on a test of this sort students who had both studied a play and seen it would have an advantage over those who had merely seen it. But in the XNOT case the scores of the after classes, which had had no classroom work connected with the play, were very significantly higher than those of the before classes, which had been prepared for the play. The means for the after and before levels were 36.52 and 34.17, respectively. This would seem like a confirmation of the wisdom of the actors' contention that students should attend the performance without preparation, in that the students who were unprepared scored better even on a test of knowledge, something which the English teachers value highly. Even the fact that the prepared

and unprepared classes were indistinguishable in regard to scores on a test of knowledge about the second play might tend to support the actors' preferences.

Additional analyses yielded a significant NOQ-NOT contrast (XNOQ − YNOQ + XNOT − YNOT; $F_{1,32} = 6.79$; $P < .01$) and a significant NOQNOTXY contrast (XNOQ − YNOQ − XNOT + YNOT; $F_{1,32} = 5.19$; $P < .03$), which may be interpreted as indicating that (1) the NOQ and NOT tests were differentially affected in the two blocks, and/or (2) that the X and Y forms of the tests are not equivalent. Still, the most parsimonious explanation of the significant within-category differences is that involving between-level differences on the XNOT test—that the students who saw *Red Roses for Me* without classroom preparation knew more about the play than those students who were prepared prior to the performance.

Hypothesis 4: The Content of the Classroom Treatment

The content factor had significant effects on scores on the quotations test of knowledge ($F_{1,32} = 4.23$; $P < .05$) and the thematic-understanding test ($F_{1,32} = 4.11$; $P < .05$). The types of learnings measured by these tests were among those highly valued by English teachers. The means by levels of the content factor were these:

	Mean Scores	
Level of Content	*NOQ*	*PHI*
Related to play	67.03	26.92
Specific to play	71.69	28.62

In both cases the classes studying materials specific to the play being performed had higher scores, which is what the educators predicted. However, the differences attributable to levels of content are few, and not large in absolute terms. It must be considered that the students who studied related materials learned things about drama and about the related plays that the students at the specific level did not learn, so it is not certain which group should be considered to have the net advantage.

When the categories of tests were considered, significant or near-significant effects were found in the knowledge ($F_{4,29} = 4.58$; $P < .01$), philosophical-insights ($F_{2,31} = 3.56$; $P < .04$), and

desirable-attitudes-and-behaviors ($F_{8,27} = 2.36$; $P < .055$) categories.

Analyses of the individual tests within the knowledge category yielded these results:

Test	$F_{1,32}$	*P*
XNOQ	2.48	0.12
XNOT	4.74	0.04
YNOQ	4.10	0.05
YNOT	4.52	0.04

For the three tests on which there were significant differences, the mean scores were:

Level of Content	*Mean Scores*		
	XNOT	*YNOQ*	*YNOT*
Related to play	27.08	33.20	30.59
Specific to play	29.00	34.84	27.55

A somewhat simpler accounting for the effect within the category may be given in terms of between-block differences and test × block interactions. Both the X-Y contrast (XNOQ − YNOQ + XNOT − YNOT) and the NOQNOTXY contrast (XNOQ − YNOQ − XNOT + YNOT) were significant (respectively, $F_{1,32} = 6.15$; $P < .02$, and $F_{1,32} = 8.18$; $P < .01$). This indicates that the effect within the knowledge category was significant because the tests were differentially affected on the two occasions—especially the true-false tests with the higher scores on the XNOT test being associated with the specific level and the higher YNOT scores being associated with the related level—and because the scores on both knowledge tests were higher in the second block than in the first. Since it seems clear that the X and Y forms of the knowledge tests may not have been equivalent—these tests were play-specific—it cannot be determined to what extent the differences are artifactual and to what extent they are due to sequence effects and differences between the plays and/or productions.

Within the philosophical-insights category, which consists of only the XPHI and YPHI tests whose summed scores have already been reported, perhaps the best explanation of the significant effect is that the overall level of the means was signifi-

cantly higher at the specific level of the content factor, a finding favoring the English teachers' position.

Within the desirable-attitudes-and-behaviors category, between-level differences were significant on the XBEH test ($F_{1,32} = 4.76$; $P < .04$), with the specific level yielding the higher mean (52.41 compared to 51.18). But the general level of the means for all tests in the category (XDAT + XBEH + XETQ + YDAT + YBEH + YETQ) were also significantly higher at the specific level ($F_{1,32} = 4.96$; $P < .03$), and since the content factor is rather tenuously related to the XBEH test considered by itself, probably the best explanation of the significant effect is that subjects who studied the specific play scored higher on all the tests in the desirable-attitudes-and-behaviors category—another finding favoring those who advocate studying the specific play.

Hypothesis 5: Interaction of the Intensity of the Study of the Background and the Intensity of the Study of the Text

There were three significant effects of the background × text interaction: on scores on the true-false knowledge test ($F_{1,32} = 7.74$; $P < .01$), the appreciation:attitudes test ($F_{1,32} = 4.11$; $P < .05$), and the thematic-understanding test ($F_{1,32} = 4.89$; $P < .04$).

The mean scores for the knowledge test were as follows:

	Levels of Text	
Levels of Background	*Brief*	*Intense*
Brief	54.29	58.60
Intense	61.43	53.83

Within the knowledge category there was a significant effect ($F_{4,29} = 2.66$; $P < .05$), which may best be explained in terms of the effects of the background × text interaction on summed means (XNOQ + YNOQ + XNOT + YNOT) and on the NOQNOTXY contrast (XNOQ − YNOQ − XNOT + YNOT). The between-level differences between means were significant ($F_{1,32} + 4.76$; $P < .04$) and described the same pattern as the means on the true-false knowledge test considered by itself.

	Levels of Text	
Levels of Background	*Brief*	*Intense*
Brief	126.9	128.5
Intense	128.5	121.9

Since the main effects of both the background and text factors were nonsignificant in regard to the knowledge tests, what probably accounts for these differences is the total duration of the classroom treatment and/or the amount of material covered in the lessons. The brief and intense levels of these factors, it should be recalled, were defined in terms of amount of material covered and number of class periods used. The data suggest that maximum familiarity with the details of a play is associated with a moderate amount of study of the play. Of particular importance is the finding that the lowest knowledge scores are associated with the most intense classroom treatment. Apparently there is a point at which students become bored or overwhelmed, so that further study has negative effects.

The remarks made at the end of the preceding section on the significant NOQNOTXY interaction ($F_{1,32} = 4.53$; $P < .04$) apply here as well.

The pattern of scores on the appreciation:attitudes test was similar to that described by the knowledge scores, with the intense-intense combination yielding the lowest scores. Effects on scores within the appreciation category were nonsignificant.

	Levels of Text	
Levels of Background	*Brief*	*Intense*
Brief	191.1	191.1
Intense	191.2	187.8

On the thematic-understanding test, however, the pattern reverses itself, and the intense-intense treatment yields the highest scores. What may be involved here is the probability that the longer a class spends studying a play, the more likely it is

	Levels of Text	
Levels of Background	*Brief*	*Intense*
Brief	27.66	27.08
Intense	27.49	28.83

that there will be explicit discussion of the kinds of issues covered on the thematic-understanding test.

Hypothesis 6: Interaction of the Intensity of Study of the Background and the Timing of the Classroom Treatment

The single significant effect of this interaction was on scores on the appreciation:cognitions test ($F_{1,32} = 4.82$; $P < .04$). The mean scores at the different combinations of levels were:

	Levels of Timing	
Levels of Background	*Before*	*After*
Brief	188.9	191.4
Intense	187.4	186.6

The appreciation:cognitions test tried to describe students' convictions about the nature and power of drama and other arts. A high score might be taken as evidence of a high opinion of the role of the arts in society. The means reported above indicate that the highest scores were associated with brief study of the backgrounds following attendance at the theatre, while the lowest scores were associated with intense study of the backgrounds following the performance. The main effects of the factors were not significant, and it is not at all clear what may be the relationship between the interaction of these two factors and the property measured by the appreciation:cognitions test. The backgrounds × timing interaction had no significant effects on scores in any of the six categories of tests, and it may be best not to try to impose an interpretation upon the single significant effect.

Hypothesis 7: Interaction of Intensity of Study of Background and the Content of the Classroom Treatment

This particular interaction had no effects either upon total test scores or upon scores within categories that approached significance. That is to say, it made no distinguishable difference whether the background studied was analytical and specific to the play performed or dramatic and related to the play performed.

Hypothesis 8: Interaction of Intensity of Study of the Text and the Timing of the Classroom Treatment

In this case, as in the preceding one, there were no significant effects at all. The effects of studying a text briefly before a

performance, briefly after a performance, intensively before a performance, or intensively after a performance were not distinguishable.

Hypothesis 9: Interaction between Intensity of Study of the Text and the Content of the Classroom Treatment

The absence of any significant effects for this particular interaction is perhaps the most surprising finding in the study. It seems to have made no difference in the students' performance whether a class studied the specific play for a week or the related play for one or two periods. If what would seem on common-sense grounds the most important sorts of differences between treatments do not produce significant effects, then the inference may reasonably be drawn that the question of the best way to study a play is a much more subtle and complex question than anyone involved in the Project was prepared to suggest.

Hypothesis 10: Interaction of the Timing of the Classroom Treatment and the Content of the Classroom Treatment

On common-sense grounds, as in the preceding case, one would predict large and numerous differences in scores due to this interaction. But again there were no significant effects, and it seems to have mattered little whether students studied the specific play before attending a performance or a related play after attending a performance. What is especially noteworthy is the lack of significant effects on such content-specific tests as those of knowledge and thematic understanding.

Hypothesis 11: Interaction of Intensity of Study of the Text, Content of the Classroom Treatment, and Timing of the Classroom Treatment

As explained above, this experiment was not specifically designed to evaluate three-factor interactions. But one of the recurring suggestions made by English teachers involved a three-factor interaction, namely, that students should intensively study (B) the text of the play (D) before attending the performance (C). We therefore had a reason for preferring the BCD interaction as an explanation of any observed significant effects over

the AE interaction with which it was aliased. But as it turned out the BCD interaction had no significant effects upon total test scores, although the effects approached significance in the case of the thematic-understanding test, the one measuring the property which English teachers most highly valued ($F_{1,32} = 3.66$; $P < .07$).

However, when the tests are grouped into categories there are two significant effects; and it so happens that these two categories are the ones corresponding to the sets of objectives that English teachers valued most highly: knowledge ($F_{4,29} = 2.86$; $P < .04$) and philosophical insights ($F_{2,31} = 3.82$; $P < .03$).

Considering the tests within these categories, differences between the different combinations of levels were significant only for the XNOT test ($F_{1,32} = 8.69$; $P < .01$) and the XPHI test ($F_{1,32} = 5.06$; $P < .03$). Both the NOT and PHI tests were administered immediately after the performance of the play, so that all the classes at an after level would have had no classroom treatment at all. All the scores for treatment conditions containing the after level of the timing factor may therefore be pooled and their means computed. The XNOT and XPHI means were as follows:

Levels of the Factors			*Mean Scores*	
Text	*Time*	*Content*	*XNOT*	*XPHI*
Brief	Before	Related	30.12	13.53
Brief	Before	Specific	30.00	15.25
Intense	Before	Related	26.87	15.19
Intense	Before	Specific	31.30	15.04
Mean of all after conditions			29.57	13.93

On the XNOT test the highest scores were associated with the combination of levels of the factors which describes the treatment advocated by the English teachers; on the XPHI test the situation is less clear-cut.

An alternative explanation of the significant effect in the knowledge category would be in terms of the X-Y contrast (XNOQ − YNOQ + XNOT − YNOT; $F_{1,32} = 6.15$; $P < .02$) with the first block means being higher in six of the eight cases. This would be in line with findings reported earlier which indicated that the treatment conditions preferred by English teachers seemed most often to work as predicted in connection with the first play.

The best explanation for the effect in the philosophical-insights category is probably in terms of the levels of mean scores ($F_{1,32} = 7.67$; $P < .01$) with the highest XPHI + YPHI scores being associated with an intense study of the specific play before the performance ($\bar{X} = 29.29$) and the lowest with an intensive study of a related play before the performance ($\bar{X} = 24.88$).

These findings tend to support the observation that each group involved preferred the combination of levels of the factors which experience indicated would maximize student gains on the tests of objectives most highly valued by the particular group.

Other Findings

Two of the tests in the interpretive-skills category have not yet been mentioned. As explained earlier, tests of interpretive skills (INT) and judgment of quality (JUD) had been written originally so that the scores could be used as covariates. We had figured that student responses on the dependent measures would probably be affected by the critical and evaluative skills students brought to the experiment. Analyses of the data from the first replication showed that once adjustments had been made to take account of variation due to verbal intelligence and prior theatre experience, scores on the INT and JUD tests accounted for very little additional variation. So it was decided to use the tests as dependent measures during the second replication of the experiment.

Used as dependent measures these tests measured transfer from the experimental treatments to performance on critical and judgmental tasks not specific either to drama or to the plays

YJUD Scores

Levels of Background	*Levels of Content*	
	Related	*Specific*
Brief	24.25	25.64
Intense	25.08	22.97

Levels of Text	*Levels of Content*	
	Related	*Specific*
Brief	24.32	25.14
Intense	25.01	23.47

that were studied. Each of the hypotheses were evaluated in regard to each of the tests and only two significant effects were found, both involving scores on the YJUD test. YJUD scores were significantly affected by the background-content interaction ($F_{1,32} = 6.06$; $P < .02$) and by the text-content interaction ($F_{1,32} = 4.00$; $P < .05$). In each case, as shown in the foregoing tables, the lowest score was obtained at the intense-specific combination of levels. As in similar cases reported earlier, such a result suggests that there is a point at which continued study becomes counterproductive.

CHAPTER EIGHT

CONCLUSIONS

Summary of Significant Effects

Within the affective-response category involvement scores seem not to have been affected by classroom treatments, while liking scores were affected differently by the timing of the classroom instruction, depending upon the play being performed.

In the knowledge category the lowest scores on all tests were associated with the most intensive classroom treatments, but there was possibly an interaction between knowledge scores and the plays being performed. The highest scores on knowledge tests were also associated with an intensive study of the text before the performance, a finding not necessarily in contradiction of the earlier finding that an intensive study of the background plus an intensive study of the text produced the lowest knowledge scores.

Within the philosophical-insights category the higher scores were associated with study of the specific play, with the intense study of both background and text, and with the intensive study of the specific play before the performance.

Within the appreciation category the lowest appreciation: attitudes scores were associated with the most intense classroom treatments and the lowest appreciation:cognitions scores were associated with intense study of the background and with intense study of the text.

Within the desirable-attitudes-and-behaviors category, higher scores on the desirable-attitudes test were associated with brief study of the background, but there were no other significant effects.

Comments

In general the relatively few effects which attained significance confirm the supposition that the English teachers preferred those arrangements which yielded the highest scores on the cognitive tasks they most highly valued. (The intensive-

95

intensive treatments which depressed knowledge scores were probably too extreme to expect they would be advocated by English teachers in general. Similarly, the actors preferred the arrangements that maximized scores in the areas of appreciation and affective response with which they were most concerned. Although each group greatly overestimated the importance of the factors, each seems to have predicted with some accuracy the effects of the factors upon student performance in the cognitive and affective areas. The case is still unsettled in the areas of attitudes and behaviors.

Further interpretations of these significant findings have already been presented and will not be repeated here. What will be repeated is that the overall impression created by the small number of significant effects is that the factors which figured in disputes about how students should be prepared for the theatre are not in themselves as important as had been thought.

Perhaps the most plausible explanation for the pattern of a scarcity of significant effects of factors which everyone agreed were important is this: The students' experiences in the theatre acted so powerfully to raise mean scores on all the dependent measures that the additional increases, or decreases, that could be effected by manipulation of the classroom treatment variables were too small in most cases to distinguish between groups of students who shared the theatre experience in common. In other words, the students may have learned about all they could learn, within the allotted span of time, from the theatrical performance itself, so that the classroom treatments taking place in conjunction with the performance were largely redundant.

The missing half of the 2^{5-1} design used in this study (Table 4) would enable one to evaluate the effects of the independent variables apart from the performances of the plays. If desired, the design could be further simplified to a 2^{5-2} design by dispensing with the distinction between the before and after levels of the timing variable. Or, alternatively, the entire 2^5 design could be executed with half the subjects attending the theatre and half not attending.

Be that as it may, the results of the present experiment do not support the positions taken either by educators or theatre people about the effects of different classroom practices as clearly as either group might have wished. Each group, however, may

take comfort from particular findings and each may care to take thought about what seems to be the relative impotence of classroom instruction to either inhibit or facilitate short-range student behaviors of the sorts measured in this study.

From the general reader's point of view, two facts should help to demonstrate that common sense, instinct, experience, and professional judgment are not necessarily good substitutes for objective, empirical evidence: (1) that different groups of experienced professionals could predict different effects for factors they agreed were important and (2) that in most cases it could not be demonstrated experimentally that these purportedly important factors had large or consistent effects in any direction. And these same facts should underscore the need for researchers to eschew techniques which are incapable of providing us with the empirical evidence which we need.

APPENDICES

APPENDIX A

DISTRIBUTION OF TEST ITEMS OVER FORMS

The various dependent measures in this experiment were distributed over three instruments, ignoring the six pretest-posttest items given to a sample of the classes. On each test were some informational items to which all students responded. These common items appeared on all ten forms of each instrument while only ten percent of the items from each of the other tests appeared on any one form. To facilitate the coding of responses and to reduce interference between similar items, items sampled from any particular test were assigned to predetermined and well-separated positions on the instruments. Table 19 shows how the items from the tests were distributed over the instruments, and the code designations in the left-hand column of the sample instruments that follow identify the test from which each item was sampled.

The first instrument was the pretest. It was given sometime before the start of the experiment and its major purpose was to get scores on the variables we planned to use as covariates: verbal intelligence, prior theatre experience, interpretive skills, and literary judgment. The other two instruments were the post-lesson test and the postperformance test. These tests differed between replications only to the extent that some of the test items were play specific. The order and number of the items on each instrument were the same for both replications. The postlesson test was administered at the end of the classroom study of the play, so that classes at the before level of the timing factor had studied but had not seen the play, while those at the after level had seen and discussed the play as well as studied it. This enabled us to compare lesson-only with lesson-plus-performance effects on certain tests. The postperformance test was administered during the first English class following attendance at the theatre. In this case, therefore, the classes at the before level had studied before attending the play, while those at the after level had attended the play without prepara-

tion. In this way one-half of the experimental classes served as a control group in regard to the timing factor.

Table 19

Distribution of Test Items over the Three Instruments

	Name of Instrument		
Name of Test	*Pretest*	*Post-lesson Test*	*Post-performance Test*
Verbal intelligence	X		
Prior theatre experience	X		
Interpretive skills	X		
Literary judgment	X		
Knowledge (true-false)		X	
Philosophical understandings		X	
Involvement			X
Knowledge (quotations)			X
Appreciation: attitudes			X
Appreciation: cognitions			X
Appreciation: discrimination			X
Desirable attitudes			X
Desirable behaviors			X
Theatre etiquette (second play only)			X
Interpretive skills		X	
Literary judgment		X	

Students were asked not to sign their names to any of the tests. It was not necessary to identify individual students in order to compute class means, and one of the informational items on each test enabled us to identify and discard the responses of students who had not attended a play. The decision to keep student responses anonymous was made in hopes of increasing the chances that students would tell us what they thought, rather than what they figured we wanted to hear. In order to gain this advantage we had to sacrifice the opportunity to refine our measurements by eliminating the responses of students who had been absent during all or most of the classroom treatment.

EXHIBIT 1: SAMPLE PRETEST

FORM 3

PreT

ANSWER SHEET

YOUR ENGLISH

*TEACHER'S NAME*_____ *DATE*_____

*YOUR SCHOOL'S NAME*_____

DIRECTIONS: Circle the letter of the answer you wish to give, according to the directions on the questionnaire. Please make sure that the number by which you place your answer on this sheet is the same as the number of the question you are answering.

1. A B C D E
2. A B C D E
3. A B C D E

EXAMPLE: A B C D (E)

4. A B C D E F
5. A B C
6. A B C

7. A B C D E
8. A B C D E
9. A B C D E
10. A B C D E
11. A B C D E
12. A B C D E

13. A B C D E
14. A B C D E

15. A B C
16. A B C

PreT Form 3
CEMREL, Inc.
ETS-1

RECORD ALL YOUR ANSWERS ON THE ANSWER SHEET BY CIRCLING THE LETTER OF THE BEST ANSWER OR ANSWERS TO EACH QUESTION

DIRECTIONS: First, fill in your school's name, your teacher's name, and the date in the spaces at the top of the answer sheet. There will be different directions for answering different groups of questions, so read these carefully as you go along. But, in all cases, you are to find, on the answer sheet, the number of the question you are answering and circle the letter that indicates the answer you wish to give.

THESE DIRECTIONS APPLY ONLY TO THE FIRST THREE QUESTIONS. *Each of these questions consists of a sentence with the first and last words left out. You are to pick out words to fill the blanks that will make the sentence true and sensible. Below each sentence are five pairs of words. The first word of the pair goes in the blank at the beginning of the sentence; the second word goes in the blank at the end. Choose the pair of words that best fills in the blanks in the sentence and circle the letter of that pair next to the number of the sentence on the answer sheet.*

EXAMPLE: . . . is to night as breakfast is to . . .

A. supper—corner
B. gentle—morning
C. door—corner
D. flow—enjoy
E. supper—morning

Only the pair of words marked E *makes sense of the sentence:* "SUPPER is to night as breakfast is to MORNING." *So you would circle* E, *as has already been done on the answer sheet.*

VIQS 1. . . . is to horse as chauffeur is to . . .

A. mane—auto
B. jockey—auto
C. stable—auto
D. mane—owner
E. mane—uniform

VIQS 2. . . . is to answer as ask is to . . .

A. question—reply
B. question—know
C. yes—reply
D. chance—reply
E. yes—know

VIQS 3. . . . is to building as designer is to . . .

A. cement—clothes
B. roof—artist
C. roof—clothes
D. architect—clothes
E. roof—modiste

THESE DIRECTIONS APPLY TO THE NEXT THREE QUESTIONS. *You may give more than one answer to question 4, but only one answer to questions 5 and 6.*

PREX 4. Have you ever participated in putting on a play for an audience? If you have, circle the letter on the answer sheet that refers to each type of work you have done.

A. I have acted a major part
B. I have acted a minor part
C. I have been in a singing or dancing chorus
D. I have worked on scenery, make-up, or other backstage jobs
E. I have worked as a ticket-taker or usher at a play
F. I have never done any work on a play

PREX 5. Have you ever seen a live play in a theatre?

A. Yes, I have seen many plays
B. Yes, I have seen one or two plays
C. No, I have never seen a live play

PREX 6. How many plays have you read or studied in your English classes?

A. Three or more
B. One or two
C. None

THESE DIRECTIONS APPLY TO QUESTIONS 7 THROUGH 12. *In each question is a statement. Read each statement and decide how strongly you agree or disagree with it. If, for instance, you think the statement is always true, you "strongly agree" with the statement. Then circle, on the answer sheet, the letter that best indicates how you feel.*

7. I watch TV much less than I did six months ago.

A. Strongly agree
B. Agree
C. I do not know
D. Disagree
E. Strongly disagree

8. Literature is the most important part of English.

A. Strongly agree
B. Agree
C. I do not know
D. Disagree
E. Strongly disagree

9. There is no reason to discuss and analyze literature; we should just read and enjoy it.

A. Strongly agree
B. Agree
C. I do not know

D. Disagree
E. Strongly disagree

10. The most important thing about literature is that it tells us how to behave morally.

A. Strongly agree
B. Agree
C. I do not know
D. Disagree
E. Strongly disagree

11. I can understand literature better if I read it aloud or act it out.

A. Strongly agree
B. Agree
C. I do not know
D. Disagree
E. Strongly disagree

12. I read much more now than I did six months ago.

A. Strongly agree
B. Agree
C. I do not know
D. Disagree
E. Strongly disagree

Read the poem below and then read the questions about it. Choose the best answer to each question, referring back to the poem as often as necessary. Circle the letter of the best answer to each question on the answer sheet.

We shall come tomorrow morning,
who were not to have her love,
We shall bring no face of envy
but a gift of praise and lilies
To the stately ceremonial
we are not the heroes of.

Let the sisters now attend her,
who are red-eyed, who are wroth;
They were younger, she was finer,
for they wearied of the waiting
And they married them to merchants,
being unbelievers both.

I was dapper when I dangled
in my pepper-and-salt;
We were only local beauties,
and we beautifully trusted
If the proud one had to tarry
we would have her by default.

But right across her threshold
has her Grizzled Baron come;
Let them wrap her as a princess,
who'd go softly down a stairway
And seal her to the stranger
for his castle in the gloom.[1]

XINT 13. Who is "the stranger" in the last line of the poem?

A. The Grizzled Baron
B. The narrator
C. Death
D. Someone from far away
E. The reader

XINT 14. Which of the following is the best statement about the rhythm of the poem?

A. It varies from stanza to stanza.
B. It is solemn and slow-moving.
C. It contrasts with the subject matter of the poem.
D. It is very lively.
E. It is very regular.

XJUD 15. Below are three versions of the same poem. Read the three versions carefully. Decide which version you like best, then circle on the answer sheet the letter that identifies that version.

A.

there are two
kinds of human
beings
first those
who could reveal
to you the secrets
of the universe but
not impress you
with the importance
of the secrets
and secondly
people who can
tell you that
they have
purchased
ten cents worth
of something
and make you
thrill and vibrate
with intelligence

B.

there are two
kinds of human
beings in the world
so my observation
has told me
namely and to wit
as follows
firstly
those who
even though they
were to reveal
the secrets of the universe
to you would fail
to impress you
with any sense
of the importance
of the news
and secondly
those who could
comunicate to you
that they had
just purchased
ten cents worth
of paper napkins
and make you
thrill and vibrate
with the intelligence

C.

there are two
kinds of human
beings in the world
so my observation
has told me
namely and to wit
as follows
firstly
those who
even though they
were to reveal to you
they had purchased
ten cents worth
of paper napkins
would fail to
impress you
with any sense
of the importance
of the news
and secondly
those who could
communicate to you
the secrets of
the universe
and make you
thrill and vibrate
with the intelligence

XJUD 16. Now look at the three poems again. Decide which version you like least, and circle the letter of that version next to number 16 on the answer sheet.

EXHIBIT 2: SAMPLE POSTLESSON TEST
FORM 4

PLT-2

ANSWER SHEET

*YOUR ENGLISH TEACHER'S NAME*_____ *DATE*____

*YOUR SCHOOL'S NAME*_____

1. A B
2. A B C D
3. A B C
4. A B C
5. A B C
6. A B C D E
7. A B C D E
8. A B C D E
9. A B C D E
10. A B C
11. A B C

PLT-2
CEMREL, Inc.
Form 4

DIRECTIONS: First, fill in your school's name, your teacher's name, and today's date in the spaces at the top of the answer sheet. There are different directions for different sections of this test, so read them carefully. All your answers are to be given on the answer sheet.

To answer questions 1 and 2, circle the letter of the proper answer on the answer sheet.

1. Have you seen the Project Discovery production of *Macbeth*?

   A. Yes
   B. No

2. Have you read all or part of Shakespeare's *Macbeth*, or have you read a story version or a summary of the play? Circle the letter of the answer which best describes how familiar you are with *Macbeth*.

   A. I have read both the play *Macbeth* and a summary of it.
   B. I have read the play *Macbeth*, but no other version of it.
   C. I have not read the play itself, but I have read a summary of it.
   D. I have read neither the play nor a summary of it.

YNOQ 3. The lines below were spoken BY one of the characters in *Macbeth*. From the list below choose the name of the character who spoke the lines and circle its letter on the answer sheet.

I am one, my liege,
Whom the vile blows and buffets of the world
Hath so incens'd that I am reckless what
I do to spite the world.

   A. Macbeth
   B. One of the murderers
   C. Lady Macbeth

YNOQ 4. The following lines from *Macbeth* were spoken TO one of the major characters. From the list below choose the name of the character being spoken to. Then circle the letter of that name on the answer sheet.

MACDUFF: Despair thy charm;
And let the angel whom thou still hast serv'd

Tell thee, Macduff was from his mother's womb
Untimely ripp'd.

A. Macbeth
B. Malcolm
C. Lady Macbeth

YPHI 5. Consider everything that happens to Lady Macbeth in the play—what she does, what she experiences, and what she may have learned from all of it. Then imagine you are able to ask one question to Lady Macbeth's ghost, and you ask the question below. Which of the three suggested answers do you think would come closest to the one Lady Macbeth's ghost would give? Circle the letter on the answer sheet that corresponds to that answer.

THE QUESTION: "It has been said that there are laws of human nature, and that according to these laws everyone will act in pretty much the same way as everyone else if the circumstances are the same. Do you think this is true?"

THE ANSWERS:

A. "Yes, I think I would agree with that. Everyone does react pretty much the same way to a given event."

B. "In my experience, the statement is untrue. How one reacts to a given event depends upon what sort of a person he is. But, I might add, one sometimes doesn't know what sort of person he is until he sees how he reacts."

C. "Well, I would have to qualify that. I would say that people who are alike will act pretty much alike in a given set of circumstances. But it is not a simple question."

6. About how many hours did your English class spend in studying or discussing the Project Discovery production of Shakespeare's *Macbeth* or matters related to it? (Include in your estimate, time spent studying other plays by Shakespeare, background materials, and drama in general; also include time spent out of class doing library research assignments; but do *not* include time spent reading a play at home.) Choose the time period below in which your estimate would fall and circle its letter on the answer sheet.

   A. Two hours or less
   B. Between two and four hours
   C. Between four and six hours
   D. Between six and eight hours
   E. More than eight hours

7. Of all the time spent in your English class studying matters related to the Project Discovery production of *Macbeth*, approximately what fraction of time was devoted to having students read aloud from the play or act out scenes from it? Choose the fraction below that comes closest to your estimate of the time devoted to acting and reading and circle its number on the answer sheet.

   A. No time
   B. One-fourth of the time
   C. One-half of the time
   D. Three quarters of the time
   E. Almost all of the time

YINT Read the poem below and then read the questions about it. Choose the best answer to each question, referring back to the poem as often as necessary. Circle the letter of the best answer to each question on the answer sheet.

The wayfarer,
Perceiving the pathway to truth,

Was struck with astonishment.
It was thickly grown with weeds.
"Ha," he said,
"I see that none has passed here
In a long time."
Later he saw that each weed
Was a singular knife.
"Well," he mumbled at last,
"Doubtless there are other roads."

YINT 8. Which of the following statements best summarizes the point of the poem?

A. The way to truth is difficult.
B. Some weeds are as sharp as knives.
C. People desire the truth, but few are willing to pay its price.
D. People are always looking for easy ways out.
E. Effort is more important than achievement.

YINT 9. Why is "the pathway to truth . . . thickly grown with weeds."

A. To show that no one has traveled the road for a long time.
B. To show the pathway to truth is soft and grassy.
C. To show that the way to truth is both dangerous and little-used.
D. To show that the pathway to truth is not used much.
E. To show that truth is a very fertile soil in which everything grows well.

YJUD 10. Below are three versions of the same stanza from a poem. Read the three versions carefully. Decide which version you like best, then circle on the answer sheet the letter that identifies that version.

A.

Were all the guns, that fill the world with terror,
Were all the wealth, bestowed on politicians,

Given to cure the human mind of error,
There were not need of buying ammunitions.

B.

Were half the power, that fills the world with terror,
Were half the wealth, bestowed on camps and courts,
Given to redeem the human mind from error;
There were no need of arsenals and forts.

C.

Were half the power that fills the world with terror,
Were all the wealth that's stolen by politicians,
Used to free men from the burdens that they bear,
And to train scientists and technicians.

YJUD 11. Now look at the three stanzas again. Decide which version you like least, and circle the letter of that version next to number 11 on the answer sheet.

EXHIBIT 3: SAMPLE POSTPERFORMANCE TEST

FORM 5 PPT-1

ANSWER SHEET

*YOUR ENGLISH TEACHER'S NAME*_____ *DATE*____

*YOUR SCHOOL'S NAME*_____

DIRECTIONS: Circle the letter of the answer you wish to give, according to the directions on the questionnaire. Please make sure the number by which you place your answer on this sheet is the same as the number of the question you are answering.

1. A B
2. A B C D E
3. A B C D
4. A B C D
5. A B C D
6. A B C D
7. A B C D
8. A B C D
9. A B C D
10. A B C D
11. A B C D
12. A B C D
13. A B C D
14. A B C D
15. A B C D
16. A B C D
17. A B C D
18. A B C D
19. A B C D
20. A B C D
21. Q L C A M Z
22. T F
23. T F
24. T F
25. T F

PPT Form 5
CEMREL, Inc.
ETS-3 (1)

DIRECTIONS: First, fill in your school's name, your teacher's name, and today's date in the spaces at the top of the answer sheet. There are different directions for different sections of this test, so read them carefully. All your answers are to be given on the answer sheet.

To answer questions 1 to 3, circle the letter of the proper answer on the answer sheet.

Questions 4 to 20 are statements. Read each statement and decide how strongly you agree or disagree with it. If, for instance, you think the statement is always true, you "strongly agree" with the statement. Then circle, on the answer sheet, the letter that best indicates how you feel.

1. Have you seen the current Project Discovery play, either with your school or in the evening?
   A. Yes
   B. No

2. Which of the following words or phrases comes closest to describing your own evaluation of the play that you just saw?
   A. Excellent
   B. Pretty good
   C. Uneven, sometimes good and sometimes poor
   D. Poor
   E. Very poor

3. Did you read the play before you saw the performance of it?
   A. Yes
   B. I read more than half of it
   C. I read part, but less than half of it
   D. No

XINV 4. I like the way that a play changes my mood.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPA 5. I think the government ought not be spending money on things like theatre.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XDAT 6. Watching the characters on stage made me realize how much one's voice conveys about him.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XINV 7. On occasion while watching a play, I've wanted to warn an actor that something was about to hurt him.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPA 8. I was more affected by seeing the play than I have been by any book that I have read.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XDAT 9. I have recognized some of my friends' faults in characters in the plays I've seen.
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XINV 10. Plays can hit me as hard as real life experiences.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPA 11. Acting plays out in class is more enjoyable than just reading them at home.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XDAT 12. Seeing plays has made me more aware of how much one is judged by his personal appearance.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPC 13. Theatre is able to present both the intellectual and emotional sides of a problem.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XETQ 14. Sometimes I was annoyed when the people sitting around me didn't seem to care about what was going on on stage.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPC 15. Since I've seen plays more I think English classes have improved.

A. Strongly agree
B. Agree

C. Disagree
D. Strongly disagree

XBEH 16. Experience in dramatics makes one more self-confident.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XAPC 17. Plays are too "preachy" to be enjoyable.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XETQ 18. I enjoy seeing an actor in different parts in different plays.

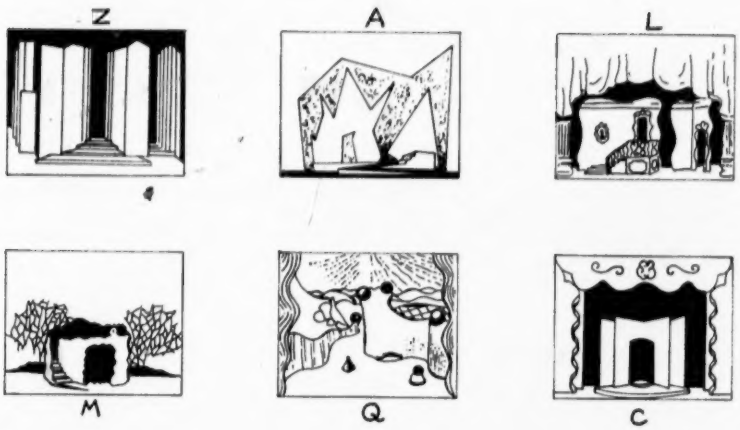
A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XBEH 19. My class seems to listen better and to be more attentive after their theatre experience.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree

XETQ 20. During the play I did not make a remark I wanted to make because I thought other students would disapprove of it.

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree



21.

DIRECTIONS. The six sketches above represent stage settings for plays. Below is a plot outline of a play. Read the plot outline and decide which of the six settings would be most appropriate for the plot. On the answer sheet, find the letter that identifies the setting you have chosen and circle it. The letters on the answer sheet are not in the same order as the pictures in most cases. Please make sure you circle the letter that you intend to circle.

THE PLOT

XADP The main characters in this play are two lonely and embittered old men, isolated from life and the world. They talk to one another and to characters who pass through about the emptiness of existence, about leaving the place where they are, and about doing something important. But at the end of the play they are still standing just where they were when the play opened, still lonely and still isolated.

The following four items are true-false questions about *Red Roses for Me*. If a statement is true, circle T on the answer sheet next to its number. If the statement is false, circle F next to the number.

XNOT 22. The two railwaymen, Dowzard and Foster, are stoned because they are Catholics.

XNOT 23. The Rector Rev. Clinton has sympathy for the Irish poor but is afraid of them.

XNOT 24. Aside from the Rector, most of the characters are tolerant of the religious beliefs of others.

XNOT 25. *Red Roses for Me* was written by Sean O'Casey.

---

[1] Copyright 1924 by Alfred A. Knopf, Inc. and renewed 1952 by John Crowe Ransom. Reprinted from *Selected Poems*, 3rd Revised Edition, by John Crowe Ransom, by permission of the publisher and Laurence Pollinger Limited, Authors' Agents.

[2] From *Archy & Mehitabel* by Don Marquis. Reprinted by permission of Doubleday & Co., Inc., New York. Copyright © 1928.

APPENDIX B

ADDITIONAL OBSERVATIONS ON THE CONDUCTING OF THE STUDY

This chapter is basically an annotated chronology of the study. It will give the reader relatively inexperienced in this sort of research a broader idea of what is involved in carrying out a study of the scope of this one; and, we hope, the remarks made upon particular arrangements and procedures will help others to profit from our experiences.

The planning for the study began in the early spring of 1968, after it had been decided that the question of how to prepare students for the plays was both important enough to justify an inquiry and well-enough defined that it could be experimentally investigated. The objectives-for-drama study was conceived of as being specifically preparatory for the experimental study. The first decision that had to be made was about the locale of the study. Rhode Island, rather than one of the other sites, was chosen primarily because the state was divided into some dozens of school districts, each of them relatively small, and our experience told us that it would be much simpler and more pleasant to carry out the study in the relatively informal atmosphere of a small school system than to try to work through a large system's bureaucracy. Another factor which recommended Rhode Island was that it seemed to us that the schools in Rhode Island had responded much more vitally and actively to the Theatre Project than had the schools in the other sites.

In early March, I visited with Donald Rock, the English department chairman at Middletown, Rhode Island, High School and at that time president of the Rhode Island Council of Teachers of English. I outlined our intentions and asked Mr. Rock to recommend to me persons who might be interested in participating in the experiment. He gave me a list of English department chairmen throughout the state whom he had reason to think would be interested. He also agreed to help us by acting as liaison between the teachers and the research staff.

In April a letter was sent to the principal of each of the schools suggested by Mr. Rock. The letter explained the proposed study and went on to state that if the principal did not express an objection we would shortly contact his English department chairman for the purpose of beginning to recruit teachers to take part in the experiment. There were no objections from principals, and shortly afterwards a memorandum was sent to the department chairmen explaining the study and asking for their assistance. Most of the chairmen recommended by Mr. Rock were indeed interested, and they sent us lists of the names of tenth-grade English teachers in their schools who had expressed an interest in the study.

Additional correspondence was sent directly to the teachers, and a late June date was selected for a planning meeting. During the weeks before this meeting the data from the objectives-for-drama study were analyzed, and an experimental design was developed in consultation with Professors David Wiley and Tom Johnson. Invitations to the planning meeting were extended to various Rhode Island school officials and to Project officers representing the schools and the theatre company.

The first meeting was a two-day affair, already discussed earlier in the report, at which the purposes of the experiment were explained, the design presented, and the teachers asked to assist in defining the experimental treatments and in writing test items. Each teacher attending the planning meeting, and the two later meetings, was given a small honorarium, as well as meals and refreshments. We think that this planning meeting played an important part in the overall success of the operation. It was immediately established that the teachers were coresearchers whose contributions were vital to the experiment.

The teachers were paid for their time as any other consultants would be. The meetings had enough of a social element that the psychological distance between researchers, teachers, and administrators was reduced. The endorsement of the experiment by respected local educators who were present at the meetings also helped immeasurably to facilitate communication and to put to rest the suspicions that are inevitably aroused when researchers come poking around in a school. The collective support of Mr. Rock, Rose Vallely, the Project Discovery Coordinator, Don Gardner, the State English Supervisor, and

Richard Cumming, the theatre company's Educational Coordinator, was especially valuable in this regard.

After the planning meeting the CEMREL staff set to work preparing the materials necessary for the study. The writing of the tests was the first order of business; next, the preparation of the instructional materials. When all of the tests had been written they were item-sampled and ten forms of each of the instruments were prepared, an elaborate job involving much shuffling of note cards and sheets of paper. One hundred fifty copies of each form of the pretest, the postlesson test, and the postperformance test were printed, collated, and stapled; and answer sheets for each of the instruments were prepared.

The instruments were assembled in sets of thirty, three copies of each of the ten forms with the forms randomly arranged. The materials for each treatment condition were collected and packed into boxes, four boxes for each of the treatments numbered 1 through 8 and three boxes for each of the treatments numbered 9 through 16. Each box contained sets of the three test instruments, but otherwise the contents of the boxes varied widely. Boxes for intensive-study of a related-text condition, for instance, contained thirty copies of O'Casey's *The Plough and the Stars*, while boxes for the brief-study of a related-text treatment contained thirty multilithed copies of a brief scene excerpted from that play. All teachers had already been given copies of the CEMREL *An Introduction to Theatre* lessons, and copies of *Red Roses for Me* were supplied from the Project offices.

In each box was a detailed description of the numbered treatment, and the treatment numbers were prominently marked on the boxes after they were sealed. The boxes were shipped to Providence in time for the meeting in early September. At this second meeting more than fifty teachers were present, but perhaps a quarter had not been at the planning meeting. Some of the original volunteers had changed their minds or had found they were not to have tenth-grade classes, while a number of new teachers coming into the schools had been interested in participating in the experiment.

The design of the experiment and the procedures that were to be followed were reviewed. Copies of the various tests were distributed and discussed, and there was a general talk session to clear up misunderstandings and to answer questions. It was

agreed that in cases of emergencies which interfered with a teacher's carrying out the treatment assigned to him the teacher should contact Mr. Rock, who would make a decision according to principles of which he was aware or forward the query to the CEMREL offices.

At the close of the meeting numbered slips of paper were handed out to the teachers, and each teacher then picked up a box marked with the same number as that on his sheet of paper. At this point the experimental study became a full time occupation for CEMREL's Rhode Island Area Coordinator, Charlotte von Breton, and her assistant, Lee McClarran. A master chart was set up, showing the treatment assigned to each teacher and the date that each school was scheduled to attend the theatre. Several days before a classroom treatment was scheduled to begin, Mrs. von Breton sent a postcard to the teachers assigned to that treatment. The card served to remind each teacher of the starting date and the details of the treatment. On the day that the last class in a particular school took the last set of tests in each replication of the experiment, Mrs. von Breton or Mrs. McClarran visited the school, picked up the sets of instruments, checked for completeness, and forwarded them to the CEMREL office.

Miss Vallely, who was in charge of scheduling school visits to the plays, cooperated in every way, sometimes rearranging schedules so that there would be ample time for teachers involved in the experiment to complete intensive treatments. Questions and problems of the sort that inevitably come up in the early stages of such an enterprise were quickly and efficiently handled by Mr. Rock and Mrs. von Breton.

The pretests were administered in mid-September, and procedures were set up for coding and key-punching the data as it was received in St. Louis. The experiment began soon after for those teachers in the before conditions for the first play, and by the time *Red Roses for Me* opened in early October things were going smoothly. The play ran through early December, and another meeting was held with the participating teachers in mid-December, at which time a preliminary report of the analyses of the available data was given and the materials for the second phase of the study were distributed. Trinity Square's *Macbeth*

opened early in January 1969, and the second phase of the experiment was under way.

The administrative arrangements remained the same, but the best laid plans of rodents and researchers gang aft agley. Or threaten to. A great many of the schools participating in the experiment had been scheduled to attend the theatre late in the run of *Macbeth* so that intense-intense treatments could be started after the Christmas holidays, thus avoiding the problem of a time lapse between classroom treatment and attendance at the theatre. It so happened that the end of the run of *Macbeth* coincided with the worst snowstorm in the memory of most Rhode Islanders. Traffic stopped; schools were closed. The final student performances of *Macbeth* were canceled, because of snow conditions and the promise of even more snow. Seventeen of our experimental classes had been scheduled to attend these canceled performances. When the situation was explained to the Project officials and the theatre management, a special performance was arranged to accommodate the experimental classes—an act of generosity clearly beyond the call of duty—and even the weatherman cooperated by being wrong about the additional snow.

Once this crisis was surmounted the rest was easy. The last experimental treatments were completed in March, and the post-tests were given to a sample of students in April. At this point we learned something of great practical value. If one wishes to use the results of a study at once as the basis on which to make decisions or plan programs, he should not get too clever for his own good. We had anticipated being able to report on the study by June or July 1969. But no available computer program was ideally suited for the data we needed analyzed. Recognizing and resolving such problems took time. The initial preparation of the data—responses of classes of varying sizes on ten forms of each of five different tests—was not always straightforward, and several repetitions of an operation were often required to assure correct results. Further, we were using a complex program for the first time to analyze data collected under an experimental plan which was novel to us. This presented us with manifold opportunities for error, and we took advantage of most of them. Each repetition and each correction of an error took more time, and the delays in giving out

the reports we had promised eventually became embarrassing. The moral is, even if you think you know precisely how you are going to get your data analyzed, be as pessimistic as possible in setting deadlines for your reporting. Something is always sure to go wrong. We were fortunate that no crucial decisions were waiting on our report; if they had been, the quite common sorts of delays we encountered, but had not adequately allowed for, might have had serious practical consequences.

In the section on factorial designs in his *Foundations of Behavioral Research*, Kerlinger notes that "four factor factorial designs . . . seem to be rare in educational research," presumably because of the difficulties inherent in manipulating so many factors (p. 327). The study that has just been reported was a five-factor fractional factorial experiment in two replications. And it worked as planned although the experimenters were most of the time a thousand miles away from the site of the experiment. We would, thinking back on it, attribute the smooth execution of an unusually complex study to the following circumstances.

1. Having had prior experience with studies in which the researchers worked through the school administration exclusively and in which the required number of teachers were more or less impressed into service by the principal, we think that it was of the most vital importance that the following things were true of this study:

    a. The teachers who took part in the study were located by working through, first, the local professional organization and the Project officials, and then the English department chairman in each building. The only contact with the school administrations was the initial one seeking permission to involve a certain number of teachers in a rather disruptive experiment.

    b. The teachers who participated in the study were volunteers who were presumably motivated in part by the fact that they perceived the problem at issue in the study as of immediate importance to themselves and their students.

c. The teachers were involved from the start as co-equal researchers and consultants and were paid and treated as professionals, not as troops to be maneuvered about. Since the teachers were experimenters themselves, rather than subjects in the experiment, each was willing conscientiously to carry out the classroom procedures he had drawn, even when his own professional judgment would have dictated quite different procedures.

d. There was frequent contact and consultation between teachers, members of the research staff, and Project officials.

2. The study was adequately financed. This meant that consultants could be brought in as needed and that the research staff was large enough to provide the necessary day-to-day administrative overseeing of the experiment and was varied enough in its talents that each part of the study was carried out by someone who knew what he was doing.

APPENDIX C

SPECIFICATIONS OF TREATMENTS AND MATERIALS FOR DESIGN CONDITION #4

First Play: O'Casey's *Red Roses for Me*

Condition #4 is specified in the design for the study as consisting of the following combination of variables for the first play.

1. Intensive, related background
2. Intensive, related text
3. Study after attending the performance of *Red Roses for Me*

The definitions of treatment variables 1 and 2 in condition #4 are summarized below for your convenience.

Intensive study of related background

A. The study will take 4–7 periods. It should follow and be separate from the general discussion of the performance which is a common part of all the treatments.

B. The subject matter will be the general orientation to drama provided in CEMREL's *Introduction to Theatre* (an edited version of Volume 1).

Intensive study of related text

A. The study will take 4–7 periods.

B. The subject matter of this study should be one of the plays in the book *Three Plays by Sean O'Casey*, which will be supplied by CEMREL. We strongly recommend *The Plough and the Stars*. The emphasis in this study should be on the dramatic elements which have been stressed in the intensive-related-background lessons. The students should act out rep-

resentative scenes in a manner similar to that prescribed for "The Marriage Proposal" in the CEMREL booklet.

Second Play: Shakespeare's *Macbeth*

Condition #4 is specified in the design for the study as consisting of the following combination of variables for the second play.

1. Brief, play-specific background
2. Brief, play-specific text
3. Study before attending the performance of *Macbeth*

Fuller definitions of these treatment variables will be forwarded to you later in the fall.

APPENDIX D

SUMMARY TABLES OF F-RATIOS AND SIGNIFICANCE LEVELS FOR TOTAL SCORES AND CATEGORIES UNDER ALL HYPOTHESES

The twenty-two summary tables in this appendix are arranged and numbered by hypotheses in the same order used in the chapter presenting the results of the experiment. For each hypothesis there are two summary tables, A and B. The A table summarizes the effects of a particular independent variable upon total scores of each of the eleven tests administered in connection with both plays. The B table summarizes the effects of that same independent variable upon scores within the six categories of tests; the F-ratio given in each case in the B tables is that for the test of equality of mean vectors. At the top of each of the tables the hypothesis being tested is stated in its null form.

Summary Table 1A
Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 1

Null Hypothesis Being Tested: There is no difference between total scores as a function of the intensity of the study of the background.

Dependent Variable	*Code Designation*	$F_{1,n}$	*P*
Liking	LIK	0.88	0.36
Involvement	INV	0.86	0.36
Knowledge (quotations)	NOQ	1.19	0.28
Knowledge (true-false)	NOT	0.00	1.00
Appreciation: attitudes	APA	0.27	0.61
Appreciation: cognitions	APC	3.09	0.09
Appreciation: discrimination	ADP	0.60	0.44
Desirable attitudes	DAT	3.62	0.07
Desirable behaviors	BEH	0.31	0.58
Theatre etiquette	ETQ	0.05	0.83
Thematic understanding	PHI	1.25	0.27

Multivariate test of equality of mean vectors: $F_{1,22} = 1.52$; $P < 0.19$

Summary Table 1B
Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 1

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the intensity of the study of background.

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	0.99	4,29	0.42
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	0.82	4,29	0.52
3. Interpretive skills	XINT, YINT, XJUD, YJUD	0.90	4,29	0.47
4. Philosophical insights	XPHI, YPHI	0.29	2,31	0.74
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.02	6,27	0.43
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	1.48	6,27	0.22

Summary Table 2A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 2

Null Hypothesis Being Tested: There is no difference between total score as a function of the intensity of the study of the text.

Dependent Variable	*Code Designation*	$F_{1,\infty}$	*P*
Liking	LIK	0.25	0.62
Involvement	INV	2.93	0.10
Knowledge (quotations)	NOQ	1.04	0.32
Knowledge (true-false)	NOT	1.27	0.27
Appreciation: attitudes	APA	5.77	0.02
Appreciation: cognitions	APC	3.27	0.08
Appreciation: discrimination	ADP	0.05	0.83
Desirable attitudes	DAT	0.21	0.65
Desirable behaviors	BEH	2.47	0.13
Theatre etiquette	ETQ	0.06	0.81
Thematic understanding	PHI	0.07	0.79

Multivariate test of equality of mean vectors: $F_{1,\infty} = 1.48$; $P < 0.21$

Summary Table 2B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 2

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the intensity of the study of the text.

Category	*Dependent Measures Within Each Category*	*F =*	*df*	*P <*
1. Affective response	XLIK, YLIK, XINV, YINV	1.49	4,29	0.22
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	0.49	4,29	0.74
3. Interpretive skills	XINT, YINT, XJUD, YJUD	0.43	4,29	0.78
4. Philosophical insights	XPHI, YPHI	0.54	2,31	0.58
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.06	6,27	0.41
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.46	6,27	0.83

Summary Table 3A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 3

Null Hypothesis Being Tested: There is no difference between total scores as a function of the timing *of the classroom treatment.*

Dependent Variable	*Code Designation*	$F_{1,\infty}$	P
Liking	LIK	0.11	0.75
Involvement	INV	0.47	0.50
Knowledge (quotations)	NOQ	1.23	0.28
Knowledge (true-false)	NOT	2.18	0.15
Appreciation: attitudes	APA	2.06	0.16
Appreciation: cognitions	APC	0.01	0.93
Appreciation: discrimination	ADP	0.19	0.67
Desirable attitudes	DAT	2.34	0.14
Desirable behaviors	BEH	0.62	0.44
Theatre etiquette	ETQ	2.69	0.11
Thematic understanding	PHI	0.41	0.53

Multivariate test of equality of mean vectors: $F_{1,\infty} = 1.13$; $P < 0.39$

Summary Table 3B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 3

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the timing *of the classroom treatment.*

Category	*Dependent Measures Within Each Category*	$F =$	*df*	$P <$
1. Affective response	XLIK, YLIK, XINV, YINV	3.07	4,29	0.03
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	3.85	4,29	0.01
3. Interpretive skills	XINT, YINT, XJUD, YJUD	2.28	4,29	0.08
4. Philosophical insights	XPHI, YPHI	0.66	2,31	0.52
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.14	6,27	0.37
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	2.11	6,27	0.09

Summary Table 4A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 4

Null Hypothesis Being Tested: There is no difference between total scores as a function of the content of the classroom treatment.

Dependent Variable	*Code Designation*	$F_{1,32}$	*P*
Liking	LIK	0.40	0.53
Involvement	INV	0.48	0.49
Knowledge (quotations)	NOQ	4.23	0.05
Knowledge (true-false)	NOT	0.32	0.58
Appreciation: attitudes	APA	1.94	0.18
Appreciation: cognitions	APC	3.19	0.09
Appreciation: discrimination	ADP	0.16	0.69
Desirable attitudes	DAT	0.00	1.00
Desirable behaviors	BEH	0.64	0.43
Theatre etiquette	ETQ	0.11	0.75
Thematic understanding	PHI	4.11	0.05

Multivariate test of equality of mean vectors: $F_{1,32} = 1.51$; $P < 0.20$

Summary Table 4B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 4

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the content of the classroom treatment.

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	2.16	4,29	0.10
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	4.58	4,29	0.01
3. Interpretive skills	XINT, YINT, XJUD, YJUD	0.40	4,29	0.80
4. Philosophical insights	XPHI, YPHI	3.56	2,31	0.04
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.48	6,27	0.22
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	2.36	6,27	0.06

Summary Table 5A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 5

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the background *and the intensity of the study of the* text.

Dependent Variable	*Code Designation*	$F_{1,\infty}$	*P*
Liking	LIK	0.16	0.70
Involvement	INV	0.05	0.82
Knowledge (quotations)	NOQ	0.17	0.69
Knowledge (true-false)	NOT	7.74	0.01
Appreciation: attitudes	APA	4.11	0.05
Appreciation: cognitions	APC	0.45	0.51
Appreciation: discrimination	ADP	0.15	0.70
Desirable attitudes	DAT	0.34	0.57
Desirable behaviors	BEH	3.30	0.08
Theatre etiquette	ETQ	0.52	0.48
Thematic understanding	PHI	4.89	0.04

Multivariate test of equality of mean vectors: $F_{1,\infty} = 2.03$; $P < 0.08$

Summary Table 5B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 5

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of background *and the intensity of the study of the* text.

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	0.30	4,29	0.88
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	2.66	4,29	0.05
3. Interpretive skills	XINT, YINT, XJUD, YJUD	1.02	4,29	0.41
4. Philosophical insights	XPHI, YPHI	1.17	2,31	0.32
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	0.31	6,27	0.93
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.87	6,27	0.53

Summary Table 6A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 6

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the background *and the* timing *of the classroom treatment.*

Dependent Variable	*Code Designation*	$F_{1,\infty}$	*P*
Liking	LIK	3.61	0.07
Involvement	INV	0.58	0.45
Knowledge (quotations)	NOQ	0.28	0.60
Knowledge (true-false)	NOT	0.09	0.77
Appreciation: attitudes	APA	0.02	0.88
Appreciation: cognitions	APC	4.82	0.04
Appreciation: discrimination	ADP	0.11	0.74
Desirable attitudes	DAT	0.01	0.92
Desirable behaviors	BEH	0.02	0.90
Theatre etiquette	ETQ	0.29	0.60
Thematic understanding	PHI	0.05	0.82

Multivariate test of equality of mean vectors: $F_{1,\infty} = 0.91$; $P < 0.56$

Summary Table 6B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 6

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of background *and the* timing *of the classroom treatment.*

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	2.36	4,29	0.08
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	0.50	4,29	0.74
3. Interpretive skills	XINT, YINT, XJUD, YJUD	1.04	4,29	0.40
4. Philosophical insights	XPHI, YPHI	0.76	2,31	0.48
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	0.86	6,27	0.54
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.34	6,27	0.90

Summary Table 7A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 7

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the background *and the* content *of the lessons.*

Dependent Variable	*Code Designation*	$F_{1,n}$	*P*
Liking	LIK	0.14	0.71
Involvement	INV	1.03	0.32
Knowledge (quotations)	NOQ	0.11	0.74
Knowledge (true-false)	NOT	3.61	0.07
Appreciation: attitudes	APA	0.18	0.67
Appreciation: cognitions	APC	0.11	0.74
Appreciation: discrimination	ADP	1.18	0.29
Desirable attitudes	DAT	0.01	0.96
Desirable behaviors	BEH	0.69	0.41
Theatre etiquette	ETQ	0.14	0.71
Thematic understanding	PHI	0.07	0.79

Multivariate test of equality of mean vectors: $F_{1,n} = 0.47$; $P < 0.92$

Summary Table 7B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 7

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of the background *and the* content *of the classroom treatment.*

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	0.54	4,29	0.70
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	1.54	4,29	0.22
3. Interpretive skills	XINT, YINT, XJUD, YJUD	1.86	4,29	0.14
4. Philosophical insights	XPHI, YPHI	0.32	2,31	0.72
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	0.80	6,27	0.58
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.22	6,27	0.96

Summary Table 8A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 8

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the text and the timing of the classroom treatment.

Dependent Variable	*Code Designation*	$F_{1,\infty}$	*P*
Liking	LIK	0.09	0.77
Involvement	INV	0.24	0.63
Knowledge (quotations)	NOQ	0.07	0.79
Knowledge (true-false)	NOT	0.03	0.86
Appreciation: attitudes	APA	1.41	0.25
Appreciation: cognitions	APC	0.61	0.44
Appreciation: discrimination	ADP	1.10	0.30
Desirable attitudes	DAT	0.02	0.91
Desirable behaviors	BEH	1.85	0.19
Theatre etiquette	ETQ	3.51	0.07
Thematic understanding	PHI	0.18	0.67

Multivariate test of equality of mean vectors: $F_{1,\infty} = 0.97$; $P < 0.51$

Summary Table 8B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 8

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of the text and the timing of the classroom treatment.

Category	*Dependent Measures Within Each Category*	*F =*	*df*	*P <*
1. Affective response	XLIK, YLIK, XINV, YINV	0.17	4,29	0.95
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	1.38	4,29	0.26
3. Interpretive skills	XINT, YINT, XJUD, YJUD	1.76	4,29	0.16
4. Philosophical insights	XPHI, YPHI	0.08	2,31	0.91
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	0.84	6,27	0.54
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	1.08	6,27	0.40

Summary Table 9A
Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 9

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the text and the content of the classroom treatment.

Dependent Variable	*Code Designation*	$F_{1,\infty}$	*P*
Liking	LIK	0.07	0.80
Involvement	INV	0.83	0.37
Knowledge (quotations)	NOQ	2.96	0.10
Knowledge (true-false)	NOT	0.26	0.61
Appreciation: attitudes	APA	1.84	0.19
Appreciation: cognitions	APC	1.15	0.29
Appreciation: discrimination	ADP	0.76	0.39
Desirable attitudes	DAT	2.99	0.10
Desirable behaviors	BEH	0.21	0.65
Theatre etiquette	ETQ	0.24	0.63
Thematic understanding	PHI	0.17	0.69

Multivariate test of equality of mean vectors: $F_{1,\infty} = 0.99$; $P < 0.50$

Summary Table 9B
Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 9

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of the text and the content of the classroom treatment.

Category	*Dependent Measures Within Each Category*	*F =*	*df*	*P <*
1. Affective response	XLIK, YLIK, XINV, YINV	0.88	4,29	0.48
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	1.46	4,29	0.24
3. Interpretive skills	XINT, YINT, XJUD, YJUD	2.22	4,29	0.09
4. Philosophical insights	XPHI, YPHI	0.38	2,31	0.68
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.25	6,27	0.31
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.52	6,27	0.78

Summary Table 10A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 10

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the timing *of the classroom treatment and the* content *of the classroom treatment.*

Dependent Variable	*Code Designation*	$F_{1,32}$	*P*
Liking	LIK	0.86	0.36
Involvement	INV	0.56	0.46
Knowledge (quotations)	NOQ	0.42	0.52
Knowledge (true-false)	NOT	0.76	0.39
Appreciation: attitudes	APA	0.03	0.86
Appreciation: cognitions	APC	4.03	0.06
Appreciation: discrimination	ADP	3.06	0.09
Desirable attitudes	DAT	3.27	0.08
Desirable behaviors	BEH	0.08	0.78
Theatre etiquette	ETQ	1.03	0.32
Thematic understanding	PHI	0.01	0.94

Multivariate test of equality of mean vectors: $F_{1,32} = 1.35$; $P < 0.26$

Summary Table 10B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 10

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the timing *and the* content *of the classroom treatment.*

Category	*Dependent Measures Within Each Category*	*F* =	*df*	*P* <
1. Affective response	XLIK, YLIK, XINV, YINV	0.50	4,29	0.73
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	0.88	4,29	0.48
3. Interpretive skills	XINT, YINT, XJUD, YJUD	0.78	4,29	0.54
4. Philosophical insights	XPHI, YPHI	1.08	2,31	0.35
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	1.74	6,27	0.15
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	1.00	6,27	0.44

Summary Table 11A

Summary of Results of Analyses of Total Scores on All Dependent Variables for Hypothesis No. 11

Null Hypothesis Being Tested: There is no difference between total scores as a function of the interaction between the intensity of the study of the text, the timing of the classroom treatment, and the content of the classroom treatment.

Dependent Variable	*Code Designation*	$F_{1,32}$	*P*
Liking	LIK	0.56	0.46
Involvement	INV	0.45	0.51
Knowledge (quotations)	NOQ	1.27	0.27
Knowledge (true-false)	NOT	1.27	0.27
Appreciation: attitudes	APA	0.39	0.54
Appreciation: cognitions	APC	2.13	0.16
Appreciation: discrimination	ADP	1.77	0.20
Desirable attitudes	DAT	0.02	0.89
Desirable behaviors	BEH	0.91	0.35
Theatre etiquette	ETQ	0.25	0.62
Thematic understanding	PHI	3.66	0.07

Multivariate test of equality of mean vectors: $F_{1,32} = 1.23$; $P < 0.33$

Summary Table 11B

Summary Results of F-ratio Tests of Equality of Mean Vectors for All Categories of Dependent Variables for Hypothesis No. 11

Null Hypothesis Being Tested: There is no difference between scores within a category as a function of the interaction between the intensity of the study of the text, the timing of the classroom treatment, and the content of the classroom treatment.

Category	*Dependent Measures Within Each Category*	*F =*	*df*	*P <*
1. Affective response	XLIK, YLIK, XINV, YINV	0.36	4,29	0.84
2. Knowledge of play	XNOQ, YNOQ, XNOT, YNOT	2.86	4,29	0.04
3. Interpretive skills	XINT, YINT, XJUD, YJUD	1.60	4,29	0.20
4. Philosophical insights	XPHI, YPHI	3.82	2,31	0.03
5. Appreciation	XAPA, YAPA, XAPC, YAPC, XADP, YADP	0.81	6,27	0.57
6. Desirable attitudes and behaviors	XDAT, YDAT, XBEH, YBEH, XETQ, YETQ	0.52	6,27	0.79

SELECTED BIBLIOGRAPHY

Campbell, Donald T., and Julian C. Stanley. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching*, edited by N.L. Gage, pp. 171–246. Chicago: Rand McNally and Co., 1963.

Edwards, Allen L. *Experimental Design in Psychological Research.* New York: Holt, Rinehart and Winston, Inc., 1968.

Evans, Bertrand. *Teaching Shakespeare in the High School.* New York: The Macmillan Co., 1966.

Furst, Edward J. *Constructing Evaluation Instruments.* New York: Longmans, Green & Co., Ltd., 1958.

Hoetker, James. *Shakespeare's* Julius Caesar: *The Initial Classroom Presentation.* St. Louis: Central Midwestern Regional Educational Laboratory (CEMREL), Inc., 1968.

————, and Alan Engelsman. *Reading a Play.* St. Louis: CEMREL, Inc., 1968.

————, and Richard Robb. "Drama in the Secondary English Class: A Study of Objectives." *Research in the Teaching of English,* Fall 1969.

————, and others. *Final Report on the Educational Laboratory Theatre Project, 1966–1970,* (four volumes). St. Louis: CEMREL, Inc., 1970.

Husek, T.R., and K. Sirontek. *Item Sampling in Educational Research.* Los Angeles: Center for the Study of Evaluation, 1967.

Kerlinger, Frederick N. *Foundations of Behavioral Research: Educational, Psychological, and Sociological Inquiry.* New York: Holt, Rinehart and Winston, Inc., 1959.

Kirk, Roger E. *Experimental Design: Procedures for the Behavioral Sciences.* Belmont, Calif.: Brooks/Cole Publishing Co., 1968.

Lord, Frederic M., and Melvin R. Novick. *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley Publishing Co., 1968.

Morison, Bradley, and Kay Fliehr. *In Search of an Audience: How an Audience Was Found for the Tyrone Guthrie Theatre.* New York: Pitman Publishing Corp., 1968.

Stephens, J.M. *The Process of Schooling: A Psychological Examination.* New York: Holt, Rinehart and Winston, Inc., 1967.

Suchman, Edward A. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs.* New York: Russell Sage Foundation, 1967.

IMPLICATIONS FOR TEACHING

On tests on which students were to judge the quality of excerpts from poems and from prose works, the scores were significantly affected by the interaction between background study and content study and by the interaction between study of the text and content study. In each case the lowest score was obtained when study was intense and specific. These results, among others, suggest that there may be an empirically establishable point at which continued study of a topic begins to have negative effects, even on the retention of purely factual knowledge of the topic under study. More generally, however—although there is some indication that both professional groups, the English teachers and the actors, were able validly to predict the effects of certain instructional arrangements on student performance—the experiment demonstrates that even those instructional variables which theory, experience, and common sense agree to be vitally important cannot necessarily be demonstrated to have, in the specific case, strikingly large effects.

IMPLICATIONS FOR RESEARCH

The discussions of the experimental arrangements make three major points about the experimental studies of English teaching:

1. Experimental research in English is most apt to be useful when the questions are practical ones and variables are definable in terms of a specific set of concrete circumstances. In other cases, naturalistic and intuitive approaches may be more fruitful.
2. The design of an experiment must be suited to the problem that is under investigation; since there are no single-variable problems in English teaching, this means that there is little or no place in the study of English teaching for experiments which evaluate the effects of one variable at a time.
3. The design and analysis of multivariate experiments is a scholarly specialty of its own, so it is incumbent upon the researcher in English to learn how to collaborate at each step in the conducting of an experiment with specialists in research design.

Finally, the study emphasizes the practical importance of conceiving of teachers with whom one may be working in an experimental study as professional colleagues and coresearchers, not as instruments for carrying out assigned tasks.